Organizational and Expressional Status of the Satellite Tagged Transcribing Genes in Gonads and Somatic Tissues of Water Buffalo *Bubalus bubalis*

ABSTRACT OF THE THESIS SUBMITTED
FOR THE AWARD OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

MARCH 2008

BY

JYOTI SRIVASTAVA

DEPARTMENT OF ZOOLOGY
FACULTY OF LIFE SCIENCES
ALIGARH MUSLIM UNIVERSITY
ALIGARH- 202 002 (U.P.)

INDIA

ABSTRACT

Repetitive sequences are dynamic components of the genome encompassing major satellites and simple sequence repeats (SSRs) which comprises minisatellites and microsatellites. Most of these SSRs are found in the non-coding genome whereas a small fraction is retained in the transcriptome which participate in gene regulation through transcription, translation, slipped strand mispairing or gene silencing. Repeat sequences are known to shrink and expand fuelling the process of copy number alteration and have been associated with tumorigenesis and several other genetic anomalies. Majority of these SSRs are evolutionarily conserved whereas others remain unique to a given genome. Their evolutionary conservation and polymorphism within and across the tissues/sex/stage/species substantiates their vital regulatory roles in the higher eukaryotes. However, the distribution and significance of SSRs within the non-coding and coding genomes, even in the best characterized organisms including human, remains unclear. To explore the organization and expression of such repeat-tagged genes, we targeted the transcriptome of water buffalo *Bubalus bubalis* as a model system. Buffalo is an important animal in agriculture, dairy and meat industries in the Indian sub-continent as India has about half of the world's buffalo population. Compared to its importance, still no information is available on the genetic makeup of this important livestock animal. Novelty also lie in the fact that buffalo genome is unexplored in terms of genes present and its association with the major satellites or SSRs.

Simple consensus repeat of a 16 nucleotide long (5' CACCTCTCCACCTGCC 3') of 33.15 repeat loci originating from the human myoglobin gene have been studied in a number of species. Similarly, the repeats of GATA and GACA sequences were identified from the Banded krait minor (*Bkm*) satellite DNA in snakes (ZW) and found to be conserved in a number of species including human, with their highest frequency on the sex chromosomes of various eukaryotes. The distribution of such important repeats in the non-coding genomes, and their organization within the mRNA

transcripts originating from somatic/gonadal tissue and spermatozoa remains largely unresolved. Owing to the tissue- and sex-specific organization of GACA, GATA and 33.15 repeats, and role of spermatozoal RNA in and post syngamy, the transcriptome fraction tagged with these repeats in somatic tissues, gonads and spermatozoa of water buffalo *Bubalus bubalis* was studied using Minisatellite/Microsatellite Associated Sequence Amplification (MASA). Moreover, the isolation and characterization of the candidate full length genes; Secreted modular calcium binding protein-1 (*Smoc*-1) and Protooncogene *c-kit* receptor *(c-kit)* were also performed from buffalo *Bubalus bubalis.* The characterization included domain organization, copy number status, *in silico* structural and functional analysis, *in-vitro* protein expression & purification, tissue & age specific transcription/translation and localization of the same onto the metaphase chromosomes & basement membrane zone. In brief, the objectives of present thesis included: **(1)** *In silico* analysis to explore the distribution of GACA, GATA and 33.15 repeats within the non-coding and coding genomes across the species. **(2)** Identification and cloning of the satellite tagged transcripts using different consensus repeat motifs and random primers following MASA with the cDNA from somatic tissues, gonads and spermatozoa in water buffalo *Bubalus bubalis.* **(3)** Sequencing and computational analysis of the MASA uncovered sequences to assess their homology status across the species, evolutionary studies and sequence organization in different tissues, if any. **(4)** Assessment of germline modulation of the MASA uncovered mRNA transcripts. **(5)** Tissue and stage specific expression for individual MASA uncovered genes/gene fragments using RNA slot blot hybridization, RT-PCR and Real Time PCR analysis. **(6)** Copy number calculation of all the MASA entrapped genes using SYBR green assays and Real Time PCR, and Chromosomal localization of candidate genes using Fluorescence *in Situ* Hybridization (FISH). **(7)** Isolation and detailed characterization of candidate genes and their *in vitro* expression studies.

Peripheral blood and tissue samples from both the sexes of water buffalo *(Bubalis bubalis)* were collected from local slaughterhouse, Delhi following strictly the guidelines of Institutional Ethical and Bio-safety

Committee. Fresh ejaculates from the buffaloes were collected from the local dairy farm. Genomic DNA was isolated from blood, tissue and semen samples using phenol: chloroform: Isoamyl-alcohol extraction method. For cross hybridization studies, DNA was also extracted from peripheral blood of cattle, sheep, goat, human, Pigeon, Pig, Baboon, Bonnet monkey, Langur, Rhesus monkey, Lion and Tiger. Lion and Tiger blood samples were procured with due approval of the competent authorities of the Government of India. RNA was isolated from blood, semen and tissue samples using Tri-X reagent as per the suppliers' specifications. The cDNA synthesis was conducted using a commercially available kit (GIBCO-BRL). MASA reactions were conducted using 6 sets of oligos based on the GACA and GATA repeats, and a 16-nucleotide long oligo for the 33.15 repeat loci. MASA uncovered a total of 148 amplicons using consensus sequence of 33.15 repeat, 332 amplicons with GACA repeat motif and 136 amplicons with GATA motif. These amplicons were cloned in to pGEMT-easy vector (Promega, USA). The resultant recombinant clones were sequenced and the sequences were deposited in the GenBank. For the characterization of buffalo *c-kit* gene, four sets of PCR primers were designed based on mRNA sequence of *Bos taurus* and *Bos primigenius*. Full length *Smoc*-1 was isolated using four sets of primers designed from the human and cattle *Smoc*-1 sequences The 5' UTR and polyadenylation signal at 3'UTR were identified using 5' & 3' RACE kits (Invitrogen, USA). For expressional analysis, RNA slot blot analysis was performed using 2 µg of total RNA from different tissues of buffalo blotted on the nylon membrane and respective clones as probes. The Northern results were further confirmed by RT-PCR using internal primers for each uncovered fragment. The evolutionary status of the uncovered genes/gene fragments was studied using them as the probes in individual southern blots with the DNA from various species. Copy number of the MASA uncovered genes and *c-kit* was calculated based on absolute quantitation assay using SYBR Green dye and Sequence Detection System-7000 (ABI, USA). For relative expression analysis, SYBR green assays were conducted for individual fragments using equal amount of cDNA from all the tissues and spermatozoa, with β-actin as an internal control. The expression level or amount of transcripts of a particular fragment was estimated by comparative Ct method

with one of the tissues as calibrator sample. Chromosome preparations were done by inoculating approximately 400 µl of whole blood in the RPMI-1640 media containing 20% fetal bovine serum. Cattle derived BAC clone Ctg9.CH240-54I18 representing full length *Smoc*-1 gene and human derived BAC clone RP11-571F15 for *Ubap*1 gene were used as probes. Probes were labeled with Fluorescein-12-dUTP using Nick Translation Kit from Vysis, (IL, USA) and detected with biotinylated anti-fluorescein antibody and FITC-avidin DCS (Vector Labs). Chromosome identification and band numbering were done through G-banding. The recombinant GST-tag-Smoc1 was transformed in to BL21 (DE3) *E.coli* and expression of the recombinant protein was induced with 1mM IPTG. Smoc-1 protein was purified using GST-tag purification resin (Clontech, USA). A rabbit was immunized with purified recombinant Smoc-1 protein using alum as an adjuvant to obtain the Anti-PSmoc1-pAb. To ensure the specificity, primary antiserum (Anti-SySmoc1-pAb) was obtained for a commercially synthesized 26 amino acid (69S to 95G) long peptide, specific to *Smoc*-1 domain, conjugated to Keyhole limpet hemocyanin (KLH). The protein was transferred onto nitrocellulose membranes, was probed with primary antibodies. Secondary detection was carried out with goat anti-rabbit IgG conjugated with HRP (Bio-rad, USA). The distribution of Smoc-1 protein in different tissues was studied on paraffin sections by indirect Immunohistochemistry using Anti-SySmoc1-pAb.

Results started with the hybridization of the GACA, GATA and 33.15 repeats with total RNA from different tissues showing discernible but differential signals indicating tagging of these repeats with several transcripts. The *in-silico* analysis of the available complete or incomplete genomes of Archea/ Eubacteria and 17 eukaryotes revealed the absence of GACA/GATA repeats in the prokaryotes demonstrating total absence of these repeats in prokaryotes and their presence in different eukaryotes studied suggested their accruement in the non-coding and coding genomes of eukaryotes. However, a gradual accumulation of these repeats was observed in the higher eukaryotes during the course of evolution. Exploration of the GACA/GATA tagged transcriptomes from lower to higher eukaryotes suggested their species-specific distribution. These repeats seem to have been acquired in

the transcriptomes with the increase in the genetic complexities in higher eukaryotes. Further, the chromosome-wise distributional studies for these repeats highlighted their concentration on the sex chromosomes of different species. Thus, the sex-chromosomal occurrence and diversity of tagged transcripts suggested the involution of the GACA/GATA repeats in functional regulation of the sex determination.

MASA using 33.15 repeat uncovered a total of 25 amplicons representing 7 different transcripts from somatic tissues, testes and ovaries and 48 amplicons comprising 12 types of transcripts from spermatozoa. GACA repeat identified a total of 57 amplicons representing 14 different types of transcripts from somatic tissues and gonads, and 104 amplicons in spermatozoa representing 26 types of transcripts. Whereas GATA repeat uncovered a total of 24 amplicons constituting 10 types of transcripts were identified from different tissues and spermatozoa, barring lung and heart which were conspicuously devoid of any amplicon. Following cloning and sequencing of all the transcripts, we observed tissue specific profile for all these genes/gene fragments. Further, the homology studies revealed that 80% of the 33.15, 65%, GACA and 10%, GATA tagged transcripts were significantly homologous with several coding genes across the species or uncharacterized BAC clones originated from cattle or human. Remaining fragments showed non-substantial or no homology with the genes present in the databank. Moreover, amongst the transcripts showing homology, only three fragments showed similarity along their entire length to the database representatives, whereas remaining ones were homologous either to 5'/3' or intervening sequences of the characterized genes. Interestingly >80% of the homologous genes were found to be involved in either signal transduction or cell-cell interaction pathways whereas remaining ~20% were implicated with several diseases. Differential transcript profiles uncovered here by different repeats may be explained either towards their various functions in somatic tissues, gonads (testis/ovary), and spermatozoa, or differential functions at various stages of development. Absence of the GATA-tagged transcripts in lung and heart is anticipated to be transcriptional quiescence of the representative genes. The homology search establishing the novel status of

approximately 40% GACA-tagged and all the GATA-tagged transcripts further corroborated their species-specific distribution.

The comparative richness of the buffalo transcriptome was analyzed for the GACA, GATA or 33.15 repeats, and found the association of relatively more number of transcripts with GACA than with GATA or 33.15 repeat. Briefly, a total of 63 different mRNA transcripts (34, GACA-tagged; 10, GATA-tagged; and 19, 33.15-tagged) representing few known and most of the novel ones were identified. Although the primates and cetartiodactyls' genomes are relatively GC poor, the GC richness of buffalo genome and transcriptome seem to be unique for its organization and thus for replication timings, genetic recombination, methylation and gene expression. There are two possible explanations for the detection of 20 of 34 GACA-tagged and 6 of 10 GATA-tagged transcripts in testis/spermatozoa. First of all, the transcripts could not be picked up in other tissues due to either polymorphic nature of the STRs or lower number of transcripts, and secondly, they are dormant in other tissues barring testis and spermatozoa. The above discussed species-specific distribution of these repeats became more interesting when these repeats picked up various mRNA transcripts in the buffalo spermatozoa as well. Although many signaling molecules and transcription factors have been reported to pass by the spermatozoan into the zygotic cytoplasm on fertilization, yet 3000-5000 transcripts remains to be characterized. The existence of the SSRs tagged transcripts in the buffalo spermatozoa is the first finding which highlights the involvements of the 33.15, GACA and GATA repeats and their tagged mRNA transcripts during the pre- and post-fertilization events. It became more significant, when all these uncovered mRNA transcripts showed faithful evolutionary conservation across thirteen different species, suggesting broader significance of these repeats in these species and may be in all the eukaryotic species.

Following, all the transcripts were analyzed for sequence polymorphisms at inter-tissue or tissue-spermatozoal levels. Of the 7 transcripts uncovered with the consensus of 33.15 repeat loci, the 846 bp one showed random nucleotide changes in the somatic tissues that did not alter

the amino acids. However, gonads (ovary and testis) showed changes at six identical places resulting in conspicuous alterations and deletions of the amino acids in the C-terminal region. The 846 bp fragment showed homology with Adenylate Kinase Like gene which is known to play an important role in reproduction. Several single nucleotide changes and INDEL polymorphisms were detected also in most of the GACA-tagged transcripts. For instance, among the transcripts detected exclusively in the different tissues, the 1.8 kb transcript depicted major alterations including insertions of 36 and 4 bp exclusively in lung and several point nucleotide changes in lung/heart or testis/ovary besides a few randomly distributed ones. The transcripts shared by spermatozoa and tissues such as 1.3 kb transcript showing homology with NFATC2 gene demonstrated the insertion of 10 bp and several single-nucleotide variations exclusively in spermatozoa. Ankyrin repeat domain-26 showed identical nucleotide sequences in both the testis and sperm, but polymorphism at several points in the ovary. This transcript was not detected in any of the somatic tissues. Next, GATA-tagged transcripts were analyzed to look for similar sequence alterations. Four out of ten transcripts evinced several single nucleotide insertions, deletions and/or substitutions at many places. DNA sequence variation can contribute to phenotypic variation by affecting the steady-level of mRNA molecules of a particular gene in a given cell or tissue. The tissue- and spermatozoa-specific sequence organizations in 30% of the repeat tagged transcripts substantiated this hypothesis.

The copy number of these repeat tagged gene/gene fragments was calculated as 1 to 65 copies per haploid genome in buffalo. The comparative expression carried out for the repeat tagged transcripts here uncovered explored the positive significant expressional variation in all the somatic/gonadal tissues and spermatozoa. Of the seven 33.15 tagged transcripts, four (AKL, LRRN6A, Spergen-3 and TCRGL) showed highest expression in testis. The expression profiles so observed suggested the potential implications of these transcripts in various testicular functions. This is an important observation because following this approach, genes expressing preferentially in gonad(s) may be easily accessed. *Smoc*-1 and TCRL genes showed highest expression in liver and spleen respectively. The 576 bp

transcript showing partial homology to TCRL-α gene seems to have immunological significance as demonstrated by its highest expression in spleen. The LRRN6A gene encoding for a transmembrane leucine-rich repeat protein involved in axonal guidance, migration, nervous system development and regeneration processes of neuronal cells also showed maximum expression in the testis.

Out of 32 GACA-tagged transcripts studied for the quantitative expressional studies, about 50% transcripts evidenced highest expression in testis, 20% in spleen/liver, and remaining 30% with uniform expression in all the tissues, when the expression was compared between somatic tissues and gonads. Further, the comparative expression of these transcripts amongst tissues and spermatozoa unveiled surprising observations such that 14 transcripts showed highest expression in the spermatozoa followed by testis and 3 exclusively in the spermatozoa. Secondly, 2 transcripts demonstrated exclusive expression in testis, 4 in liver/spleen and 9 showed consistent expression in all the sources studied. Interestingly, the highest expression of a total of 29 transcripts comprising 19 GACA- and all the 10 GATA-tagged transcripts in testis and/or spermatozoa corroborated their deep involutions in the spermatogenesis and fertilization events. The negligible or lower expression of these gene fragments in other tissues including ovary, further substantiated their potentials in the male gonad development. Studies have hypothesized the participation of GACA/GATA repeats in reproduction and heterogametic germ cell development. Present study demonstrating the testis- and spermatozoa-specific expression of majority of the GACA/GATA tagged transcripts further substantiates this hypothesis.

Chromosomal localization mapped the *Ubap*1 gene onto short arm of the metacentric chromosome 3 whereas ANKD26 onto the proximal end of short arm of the sub-metacentric chromosome 4 in water buffalo.

Proto-oncogene *c-kit* receptor is implicated with spermatogenesis, melanogenesis and hematopoiesis, and undergoes tissue/stage specific alternate splicing. The 2973 bp full length cDNA sequence was isolated from

different tissues of buffalo. Upon comparison, the *c-kit* sequences showed tissue specific nucleotide changes resulting in novel truncated peptide. These peptide lacked intra-cellular and/or transmembrane domains in all the tissues except testis. Other alternatively spliced tissue specific transcripts were also detected, which are the integral part of the open reading frame and have been reported in other mammals. Phylogenetic analysis of the sequences revealed unique tyrosine kinase domain in buffalo. Copy number calculation and expression analysis of *c-kit* established its single copy status and highest expression (137-177 folds) in testis compared to that in liver. *C-kit* expression was detected in semen samples although 10 times lesser compared to that in testis. The highest expression of *c-kit* in testis and the presence of mRNA transcript in sperms substantiate its predominant role in spermatogenesis.

Secreted modular calcium binding protein-1 (*Smoc*-1) belongs to the BM-40 family which has been implicated with tissue remodeling, angiogenesis and bone mineralization. We detected the partial *Smoc*-1 tagged with the 33.15 repeat loci in buffalo. We further cloned and characterized the full length *Smoc*-1 including its copy number status, *in-vitro* protein expression, tissue & age specific transcription/translation, chromosomal mapping and localization on to the basement membrane zone. The buffalo *Smoc*-1 was found to encode a secreted matricellular glycoprotein containing EF-hand calcium binding motifs homologous to that of the BM-40 family. This single copy gene contained 12 exons and was mapped on to the acrocentric chromosome 11. Though this gene was found to be evolutionarily conserved, the buffalo *Smoc*-1 showed conspicuous nucleotide/amino acid changes altering its secondary structure compared to that in other mammals. *In silico* analysis of the Smoc-1 proposed its glycoprotein nature with a calcium dependent conformation. Further we unveiled two transcript variants of this gene, varying in their 3' UTR lengths but both coding for identical proteins. *Smoc*-1 evinced highest expression of both the variants in liver and modest to negligible in other tissues. The relative expression of variant 02 was markedly higher compared to that of variant 01 in all the tissues examined. Moreover, the expression of *Smoc*-1, though modest during the early ages, was conspicuously enhanced after one year and remained consistently higher

during the entire life-span of buffalo with gradual increment in the expression of variant 02. Immunohistochemically, Smoc-1 was localized in the basement membrane zone and extracellular matrices of various tissues. These data added to our understanding about the tissue, age and species specific functions of the *Smoc*-1.

Present study deals with the identification and characterization of several mRNA transcripts tagged with the GACA, GATA and 33.15 repeats with the somatic as well as spermatozoal transcriptomes establishing the GACA richness of genome of water buffalo, *Bubalus bubalis*. The uncovered mRNA transcripts were found to be involved in several pathways such as signal transduction, transcription, translation, immunological activities, and sex-differentiation. In addition, the sequence polymorphisms and differential gene expression was observed suggesting the diverse functions of these repeat-tagged transcripts in different cell types, ages, stages and tissues. Moreover, the highest expression of the GACA/GATA tagged transcripts in testis and/or spermatozoa indicates their crucial roles in male gametogenesis. The detailed isolation and characterization of the full length *Smoc*-1 and *c-kit* genes were also performed to gain insight into their structural and functional organization and expressional status. MASA mediated approach seems to be highly effective for isolating a large number of mRNA transcripts which harbor the consensus of these repeats. It can be used as a basis for contemplating other repeats to establish their combined conclusive significance within and adjacent to the coding regions. The functional studies of the transcripts so uncovered will resolve the enigma of such simple sequence repeats in the mammalian genome.

Organizational and Expressional Status of the Satellite Tagged Transcribing Genes in Gonads and Somatic Tissues of Water Buffalo *Bubalus bubalis*

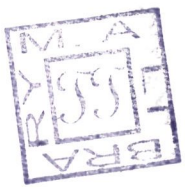
THESIS SUBMITTED TO
ALIGARH MUSLIM UNIVERSITY
FOR THE AWARD OF THE DEGREE
OF

DOCTOR OF PHILOSOPHY

MARCH 2008

BY

JYOTI SRIVASTAVA
DEPARTMENT OF ZOOLOGY
FACULTY OF LIFE SCIENCES
ALIGARH MUSLIM UNIVERSITY
ALIGARH- 202 002 (U.P.)
INDIA



FACULTY OF LIFE SCIENCES

ALIGARH MUSLIM UNIVERSITY
ALIGARH- 202 002
UTTAR PRADESH, INDIA

CERTIFICATE

This is to certify that the work embodied in this thesis entitled **"Organizational and Expressional Status of the Satellite Tagged Transcribing Genes in Gonads and Somatic Tissues of Water Buffalo *Bubalus bubalis*"** was carried out by **Jyoti Srivastava**, at the Department of Zoology, Aligarh Muslim University, Aligarh- 202 002, and the Molecular Genetics Laboratory, National Institute of Immunology, New Delhi- 110 067. This work is original and has not been submitted in part or full for any other diploma or degree of any university.

Professor Iqbal Parwez
(Supervisor)
Department of Zoology
Aligarh Muslim University
Aligarh
Uttar Pradesh-202 002
India

Dr. Sher Ali
(Co-Supervisor)
Staff Scientist and Chief
Molecular Genetics Laboratory
National Institute of Immunology
New Delhi-110 067
India

Dedicated to My Parents

ACKNOWLEDGEMENTS

The words are inadequate to express my gratitude to my esteemed supervisors, Prof. Iqbal Parwez and Dr. Sher Ali for their constant encouragement, valuable guidance and whole-hearted support at every stage of my attempt in this research program. I am beholden to Dr. Sher Ali for keeping his faith in me and my potentials. This thesis, I believe, is the outcome of his punctilious scanning of the manuscripts, patiently going through to improve them again and again.

My sincere thanks are due to Professor Absar Mustafa Khan, Head, Department of Zoology, Aligarh Muslim University, Aligarh (U.P.). I would also like to thank Professor Awadhesh K. Surolia, Director, National Institute of Immunology, for his support.

I am also grateful to Dr. Sudhir Kumar, Dr. Lalit C. Garg, Dr. Ravi Dhar and Dr. Tapan K. Chaudhuri for their generous help and suggestions. My special thanks to Professor Sita Naik and Dr. P.C. Srivastava for introducing me to this fascinating world of molecular biology. I would also like to express my thanks to Mr. Khem Singh Negi, for his technical assistance.

I wish to take the opportunity to thank the past members of MGL, Rahman, Aparna, Dr. Sebastian, Jayita and Ganesan; and present lab members, Sanjay, Safdar, Dr. Ritu, Deepali and Dr. Rana for their support. My special thanks to Avnish, Arif, Subuhi and Abuzar for their generous help. My heartful thanks to all my friends especially Smita, Ashwani, Vipin, Gaurav, Deepak, Manoj, Dr. Kanchan and Preeethy for their constant heartful wishes. I also thank Mr. Babulal for Xeroxing and binding, and all the people in the store departments and administration for their invaluable help.

I do not have adequate words to express my heartily gratitude to my parents and my fiancée Sanjay for their omnipresent support and encouragement, who were always there for me at every moment either ups or downs. Without their love and understandings, I would never have come this far.

jyoti Srivastava

JYOTI SRIVASTAVA

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

aa	amino acids
AKL	Adenylate Kinase Like
ANKD26	Ankyrin Repeat Domain-26
AP-PCR	Arbitrarily Primed Polymerase Chain Reaction
BAC	Bacterial Artificial Chromosome
BAX	BCL2-associated X protein
BM-40	Basement Membrane-40
BM-40	Basement membrane-40
BSA	Bovine Serum Albumin
C/EBPb	CCAAT/Enhancer Binding Protein b
CAT	Chloramphenicol Aceyltransferase
cDNA	Complementary DNA
CDS	cDNA Sequence
CI	Chloroform: Isoamylal alcohol
CFTR	Cystic fibrosis transmembrane conductance regulator gene
CHK1	Checkpoint Homologue Kinase-1
Ct	Cycle threshold
CUGBP1	CUG Repeat Binding Protein
DAB	Di-amino Benzene
DM	Dystrophia Myotonica
DRPLA	Dentatorubro-pallidoluysian Atrophy
E2F4	E2F Transcription Factor 4
EC	Extra-cellular
ECD	Extra-cellular Domain
ECM	Extracellular matrix
EDTA	Ethylene Diamine Tetra-acetic Acid
EPM1	Epilepsy, Progressive Myoclonic 1
FISH	Fluorescence in Situ Hybridization
FRAXA	Fragile X Syndrome
FRDA	Friedreich's Ataxia
FRDA	Fumarate Reductase
FRP	Flavin Oxidoreducatse
FS	Follistatin
FSHMD1A	Facioscapulohumeral Muscular Dystrophy 1A
FSmoc-1	Full length cDNA sequence of *Smoc*-1
GST	Glutathione S-transferase
HBGF-1	Heparin Binding Growth Factor-1
hCALM1	Human Calmodulin-1 Gene
HD	Huntington Disease

hMSH	Human mutS homologue
HRP	Horse Radish Peroxidase
ICD	Intracellular Domain
Ig	Immunoglobulin
IGFIIR	Insulin Growth Factor II Receptor
IPTG	Isopropyl Thiogalactopyranoside
ISCNDB	International System for Chromosome Nomenclature of Domestic Bovids
KI	Kinase Insert
KLH	Keyhole Limpet Hemocyanin
LRRN6A	Leucine Rich Repeat Neuronal 6A
MASA	Minisatellite/Microsatellite Associated Sequence Amplification
M-CSF	Macrophage-colony-stimulating factor
MLH3	mutL Homologue 3
MMR	Mismatch Repair
MOPS	3-(N-morpholino) propanesulphonic Acid
NFATC2	Nuclear Factor of Activated T-Cells
PAGE	Poly-Acrylamide Gel Electrophoresis
PBMCs	Peripheral blood mononuclear cells
PBS	Phosphate Buffer Saline
PCR	Polymerase Chain Reaction
PDGF	Platelet derived growth factor
PMS2	Post-meiotic Segregation Increased 2
Poly A	Polyadenylation
PSmoc-1	Partial sequence of Smoc-1
PV	Phase Variation
RACE	Random Amplification of cDNA Ends
RAPD	Random Amplified Polymorphic DNA fragment
RPMI	Roswell Park Memorial Institute Medium
RT	Room Temperature
RTK	Receptor Tyrosine Kinase
RT-PCR	Reverse Transcriptase PCR
SBMA	Spinal Bulbar muscular Atrophy
SC1	Transcription Factor-19
SCA	Spino-Cerebellar Ataxias
SDS	Sodium Dodecyl Sulphate
siRNA	Small Interfering RNA
Smoc-1	Secreted Modular Calcium Binding Protein-1
SPARC	Secreted Protein Acidic and Rich in Cysteine
SSRs	Simple Sequence Repeats
st DNA	Satellite DNAs

TBE	Tris-Borate EDTA
TBS	Tris-Buffered Saline
TCRGL	T-Cell Receptor Gamma Like
TCRL	T-Cell Receptor Like
TGFbetaIIR	Transforming Growth Factor Beta II Receptor
Tm	Melting Temperature
TMD	Trans-membrane Domain
TMEFF	Transmembrane Protein with EGF-like and Two Follistatin like Domains
TY	Thyroglobulin
Ubap1	Ubiquitin-associated protein-1
UTR	Untranslated Region
VNTRs	Variable Number of Tandem Repeats
WASF2	WAS Protein Family, Member 2

INTRODUCTION

1. INTRODUCTION AND OBJECTIVES

The eukaryotic genome contains a predominant portion (~55%) of different repetitive sequences and a small (2-3%) portion of mature transcripts (Ugarkovic, 1995; Bennett, 2000; Jasinska *et al.*, 2004). Repetitive sequences are dynamic components of the genome encompassing major satellites and simple sequence repeats (SSRs) comprising minisatellites and microsatellites (Charlesworth *et al.*, 1994; Jeffereys *et al.*, 1998). The highly polymorphic and multi-allelic SSRs (Tautz, 1989) containing short tandem iterations, are potentially involved in genome evolution by creating and maintaining genetic variability (Bennett, 2000; Toth *et al.*, 2000; Verstrepen *et al.*, 2005). Most of these SSRs are found in non-coding genome whereas a small fraction is retained in the transcriptome (Bennett, 2000; Jasinska *et al.*, 2004) which participate in gene regulation through transcription, translation, slipped strand mispairing or gene silencing (Rocha *et al.*, 2002; Li *et al.*, 2004). Repeat sequences are known to shrink and expand fuelling the process of copy number alteration (John and Ali, 1997; Nakamura *et al.*, 1987) and have been associated with tumorigenesis and several other genetic anomalies (Epplen, 1988; Kizawa *et al.*, 2005; Ross *et al.*, 2005). The expansion and contraction of the SSRs within the protein-coding sequences is proposed to modulate disease risks such as Huntington's disease, Myotonic dystrophy and fragile X Syndrome (Sutherland and Richards, 1995; Richards, 2001; Borstnik and Pumpernik, 2002; Di Prospero and Fischbeck, 2005; Dushlaine *et al.*, 2005). Majority of these SSRs are evolutionarily conserved (Robles *et al.*, 2004; Tautz, 1989) whereas others remain unique to a given genome (Ali *et al.*, 1999). Their evolutionary conservation and polymorphism within and across the tissues/sex/stage/species substantiates their vital regulatory roles in higher eukaryotes (Tautz, 1989; Ali *et al.*, 1999; Robles *et al.*, 2004). However, the distribution and significance of SSRs within the non-coding and coding genomes, even in the best characterized organisms including human, remains unclear.

To explore the organization and expression of such repeat-tagged genes, we targeted the transcriptome of water buffalo *Bubalus bubalis* as a model system. Buffalo is an important animal in agriculture, dairy and meat industries in the Indian sub-continent as India has about half of the world's buffalo population. Compared to its importance, still no information is available yet on the genetic makeup of this important livestock animal. Novelty also lied in the fact that buffalo genome is unexplored in terms of genes present and its association with the major satellites or SSRs.

Simple consensus repeat of 16 nucleotide long (5' CACCTCTCCACCTGCC 3') of 33.15 repeat loci originating from the human myoglobin gene have been studied in a number of species (Ali and Wallace, 1988; Jeffreys *et al.*, 1985; Weitzel *et al.*, 1988). The repeats of GACA and GATA sequences were identified from the Banded krait minor (*Bkm*) satellite DNA in snakes (ZW) and found to be conserved in a number of species including human, with their highest frequency on the sex chromosomes of various eukaryotes (Epplen *et al.*, 1982; Singh and Jones, 1982; Singh *et al.*, 1980; Hobza *et al.*, 2006). High condensation of these repeats in somatic cells and decondensation in germ cells during early ages of development with sex-/tissue-specific expression in higher eukaryotes supported their crucial role in sex differentiation (Singh and Jones, 1982; Singh *et al.*, 1994; Subramanian *et al.*, 2003). However, the distribution of such important repeats in the non-coding genomes, and their organization within the mRNA transcripts originating from somatic/gonadal tissue and spermatozoa remain largely unclear.

Ejaculate spermatozoa are terminally differentiated cells in which transcription and/or translation of nuclear encoded mRNAs are unlikely. Therefore, until recently, the male genome was considered to be the only cargo carried by the spermatozoa. The discovery of many soluble signaling molecules, transcription factors and structures such as centriole being introduced by spermatozoan into the zygotic cytoplasm upon fertilization has changed this perception (Saunders *et al.*, 2002; Krawetz, 2005; Miller *et al.*, 2005). Despite the transcriptionally dormant state, spermatozoa retain an entourage of transcripts, encoding transcription factors and proteins involved in signal transduction, cell proliferation, DNA

condensation, regulation of sperm motility, capacitation and acrosome reaction (Wykes *et al.*, 1997; Miller, 2000; Lambard *et al.*, 2004; Krawetz, 2005; Miller *et al.*, 2005; Ostermeier *et al.*, 2005). The delivery of such spermatozoal transcripts to ooplasm entails their potential significance during fertilization, embryogenesis and morphogenesis.

Owing to the tissue- and sex-specific organization of GACA, GATA and 33.15 repeats and role of spermatozoal RNA in and post syngamy, the transcriptome fraction tagged with these repeats in somatic/gonadal tissues and spermatozoa of water buffalo *Bubalus bubalis* using Minisatellite/Microsatellite Associated Sequence Amplification (MASA) was studied. The transcripts so uncovered were characterized in detail. Moreover, the detailed isolation and characterization of the candidate full length genes; Secreted modular calcium binding protein-1 (*Smoc*-1) and Protooncogene *c-kit* receptor *(c-kit)* were also performed from buffalo *Bubalus bubalis.* The characterization included domain organization, copy number status, *in silico* structural and functional analysis, *in-vitro* protein expression & purification, tissue & age specific transcription/translation and localization of the same onto the metaphase chromosomes & basement membrane zone. *Smoc*-1 belongs to the BM-40 family which has been implicated with tissue remodeling, angiogenesis and bone mineralization whereas *c-kit* is implicated with spermatogenesis, melanogenesis and hematopoeisis. Besides their anticipated roles in such important pathways, *Smoc*-1 and *c-kit* has been characterized only in a few mammalian species. We took advantage of their association with the repeats to characterize these genes. However, the brief objectives of present thesis included:

1. *In silico* analysis to explore the distribution of GACA, GATA and 33.15 repeats within the non-coding and coding genomes across the species.

2. Identification and cloning of the satellite tagged transcripts using different consensus repeat motifs and random primers following MASA with the cDNA from somatic tissues, gonads and spermatozoa in water buffalo *Bubalus bubalis.*

3. Sequencing and computational analysis of the MASA uncovered sequences to assess their homology status across the species, evolutionary studies and sequence organization in different tissues, if any.

4. Assessment of germline modulation of the MASA uncovered mRNA transcripts.

5. Tissue and stage specific expression for individual MASA uncovered genes/gene fragments using RNA slot blot hybridization, RT-PCR and Real Time PCR analysis.

6. Copy number calculation of all the MASA entrapped genes using SYBR green assays and Real Time PCR, and Chromosomal localization of candidate genes using Fluorescence in Situ Hybridization (FISH).

7. Isolation and detailed characterization of candidate genes and their *in vitro* expression studies.

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

2.1 *Bubalus bubalis:* An important livestock species

The family Bovidae, Sub-order-Ruminantia, Order-Artiodactyla, Sub-class-Ungulata, Class-Mammalia includes four different group of livestock animals- bovines (cattle), bubaline (buffalo), caprines (goat) and ovines (sheep). Of these, buffalo (*Bubalus*) is the indigenous breed important for its contributions in dairy, meat and agricultural industries. The world's buffalo population has been classified into two groups- the African buffalo *Syncerus* and the Asian buffalo *Bubalus.* Detailed information on the origin and precise period of domestication of the buffalo is bewildered in antiquity. However, buffalo has been domesticated in the Indian subcontinent around 5000 years ago whereas the domestication of swamp buffaloes took place independently in China about 1000 years later. India has about half of the buffalo population which is mostly of the riverine type having 50 chromosomes. The buffaloes of South East Asia and China are of the swamp type having 48 chromosomes.

The riverine buffaloes constitute about 65% of the total world buffalo population which account for 92% of the total milk and 13.9% of the meat produced every year. Riverine buffaloes, though fewer in number than cattle contribute more than 60% of the milk required for the human consumption. Inspite of being an indigenous breed as well as the most important live stock in our country, the management and breeding practices of this species have remained fallible/ unexplored for years. No systematic attempts have been made for the genetic improvement of this neglected animal species. In the past, selection criteria for buffaloes breeding have been based on phenotypic observations, blood groups, biochemical polymorphisms and mitochondrial DNA polymorphisms. There exists a need to upgrade the poor-yielding/non-descript buffaloes with germplasm from the superior breeds. Selection of superior germplasm from improved breeds for milk and meat production needs to be performed using progeny testing. The potential of this elite species needs to be explored in depth since India has dominated the world trade in export of this exotic species in the last decade. To improve this important animal,

the buffalo genome has to be explored in terms of genes present therein and their association with the regulatory elements such as repetitive DNA sequences.

2.2 Repetitive Sequences

Eukaryotic genomes are very complex and dynamic entities containing much more DNA than needed for encoding proteins, different RNAs and regulatory elements (Figure 1). Repetitive sequences form this large fraction (~50%) of the genomes that can count for this extra DNA. These repetitive elements interact with whole genome either to influence its evolution or to interact with nearby genes in several manners. However, based on the primary organization, the repetitive sequences can be classified as interspersed and tandem repeated DNA (Charlesworth *et al.*, 1994). The first class consists of sequences scattered throughout the genome, also known as transposable elements (or mobile genetic elements) because of their ability of 'jumping' to different genomic locations (transposition). These interspersed repeats have been further divided into two categories; retero-transposons and transposons, according to their mechanism of transposition, which has been studied in detail in the mammalian genomes (Capy *et al.*, 1997). These elements may silence a gene by interrupting its coding or regulatory sequences, and are thought to be responsible for gene (or exon) shuffling and duplication (Ogata *et al.*, 2000). The regulatory sequences carrying these elements can also override the normal expression pattern of the gene, provoking the alterations in expression level, timing and/or tissue specificity (McDonald, 1995, Ting *et al.*, 1992).

The second major type of repetitive DNA, tandemly repeated DNA, here discussed in detail has long presented a genomic puzzle: they are ubiquitous, yet conserved in neither content nor occurrence. These sequences are virtually present in higher copy number than the transposable elements. Tandemly repetitive sequences can be broadly categorized into three major groups which have been discussed below:

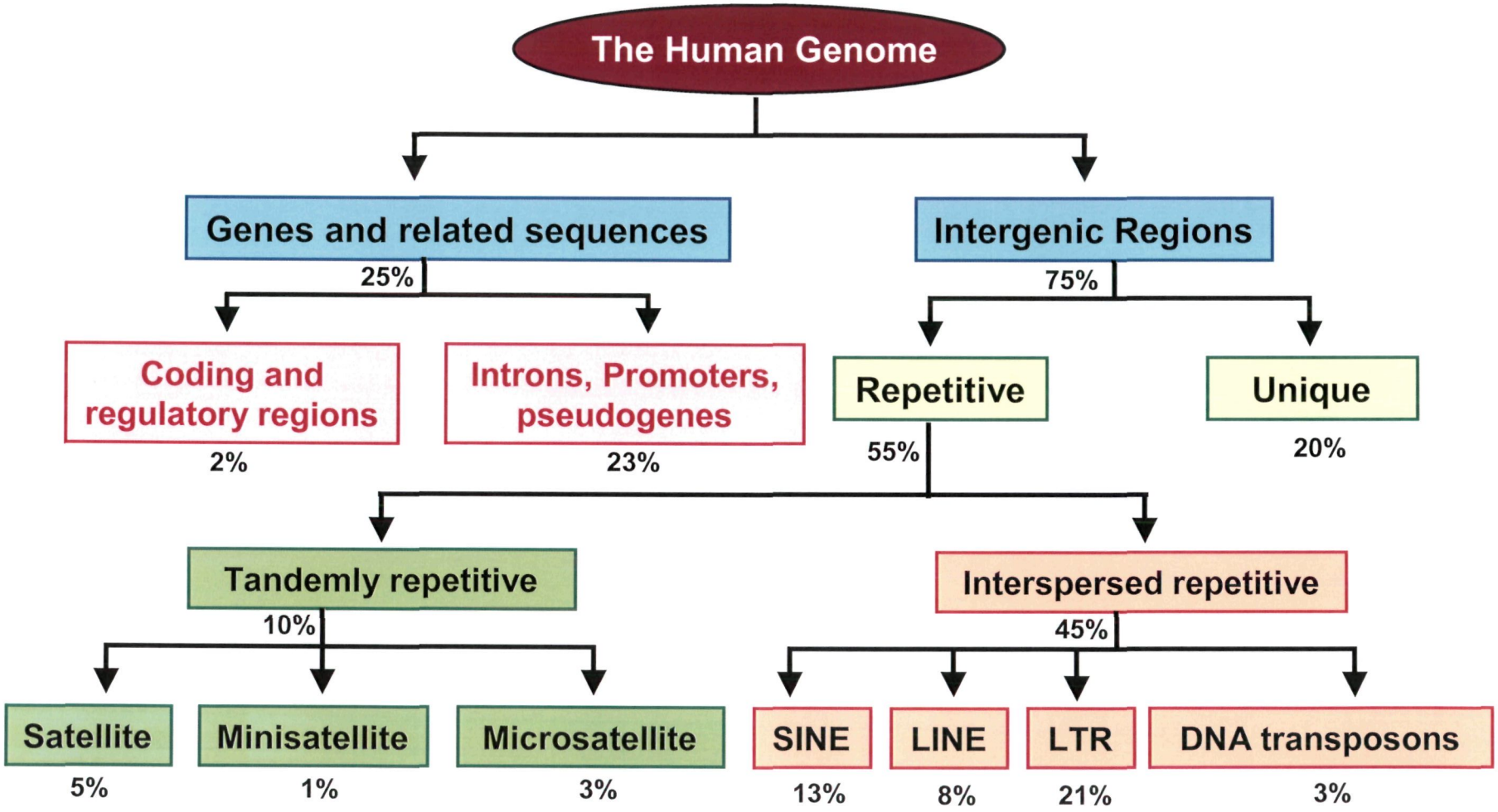


Figure 1. Composition of the human genome. The percentage shares of various functional and non-functional sequences are shown.

2.2.1 Satellites

These sequences are highly repetitive, organized in long repeats (lengths of one to several thousand base pairs), usually megabase-sized arrays representing a large portion of the genome, and are located in regions of pericentromeric and/or telomeric heterochromatin (Charlesworth *et al.*, 1994). Satellite DNAs (st DNA) are more abundant in eukaryotes and less common in prokaryotes. However, they account for the majority of genomic DNA in some species such as kangaroo, rat and beetles from the coleopteran family Tenebrionidae (Petitpierre *et al.*, 1995). Satellite DNA has the unusual property of being highly unstable, therefore, is unique to a species. There also exist various types of internal sequence organizations within the repeats. In humans, apart from st DNAs I, II, III, IV, A, B, C, the alphoid st DNAs are also described. Basic repeat units of st DNAs usually have distinct complex sequences, such as the 171-bp-long monomer of the human α-satellite, which represents a main structural element of centromeric and pericentromeric regions (Schueler *et al.*, 2001). The most abundant mouse pericentromeric γ- satellite also belongs to this group as it is composed of 234-bp monomers of specific sequence (Rudert *et al.*, 1995). However, other satellites are composed of short simple repeats, such as human satellite III with a 5 bp-long monomer, as well as many of the *Drosophila* satellites (Borstnik *et al.*, 1994).

Several type of st DNAs are subjected to the influence of gene conversion and unequal crossing over. These recombinational mechanisms are responsible for the rapid horizontal spread of mutations among monomers in a genome. These mutations are subsequently fixed in reproductive populations through the stochastic process of molecular drive (Dover, 1986). Copy number change and loss of the st DNAs from the genome are also the result of this unequal crossing over. The outcome of the recombinational mechanisms and molecular drive is a high turnover of this part of the eukaryotic genome. Therefore, st DNAs show significant rearrangements and sequence divergence as well as changes in copy number, even between closely related species (Ugarkovic and Plohl, 2002). In some cases, changes in satellite DNA profiles can be correlated

with the chromosomal evolution and could possibly influence the evolution of species.

2.2.1.1 Satellites in the non-coding regions of the genome

Earlier studies denied any function for these abundant genomic components, proclaiming them to be 'junk' or 'detritus'. Later, these satellite DNAs were found to be associated with the complex organizational features such as heterochromatic genome compartments important for proper chromosomal behavior in mitosis and meiosis (e.g. sister chromatid pairing and chromosome association) (Csink and Henikoff, 1998). Despite their structural divergence and general lack of sequence conservation across species, st DNAs were found to be major centromere-building element as has been shown in detail in *Drosophila melanogaster* (Sun *et al.*, 1997) and in humans (Schueler *et al.*, 2001). Recent studies unveiled that these sequences are associated strongly with several proteins to form unique centromeric heterochromatin and participate in centromeric condensation (Henikoff and Dalal, 2005). Satellite DNAs are not, however, a prerequisite for centromere establishment and are instead proposed to drive the adaptive evolution of specific centromeric histones (Cooper and Henikoff, 2004). Recent data indicate that the evolution of satellite DNA sequences is not only driven by molecular drive, but also influenced by selective constraints (Hall *et al.*, 2003; Mravinac *et al.*, 2005). Selective constraints on satellite sequence are probably related to their interaction with specific proteins necessary for heterochromatin formation and regulation of gene expression.

2.2.1.2 Satellite DNA transcripts and their implications

Satellites were initially thought to be of non-coding nature, which participate only in genome organization. Recently, the transcripts of st DNAs have been reported in several organisms including vertebrates, invertebrates and plants. In most species, st DNAs are temporally transcribed at particular developmental stages or are differentially expressed in some cell types, tissues or organs. Transcription from st DNA has been detected during embryogenesis in the newts *Triturus cristatus*

carnifex (Varley *et al.*, 1980) and *Notophthalmus viridescens* (Diaz *et al.*, 1981). Transcripts of an α-like satellite repeat detected during early embryogenesis in chick and zebrafish were limited to the cardiac neural crest, the head and the heart (Li and Kirby, 2003). Two types of transcripts were identified; one that corresponds to α-repeat RNA and another group of mRNAs that contain an α-like satellite sequence in the 5' and 3' untranslated regions. Mouse γ-satellite DNA is differentially expressed during development of the central nervous system, as well as in the adult liver and testis (Rudert *et al.*, 1995). Most satellite transcripts are present as polyadenylated RNA in the cytoplasm but some are found exclusively in the nucleus, such as those associated with the Y chromosome of *D. melanogaster* and *D. hydei* (Trapitz *et al.*, 1988; Bonaccorsi *et al.*, 1990). The developmental, stage- and tissue-specific expression of st DNAs in several species suggests their regulatory roles, although for most transcripts this role is still elusive and hypothetical. Taking into account the extreme sequence diversity of st DNAs and their transcripts, several sequence-specific regulatory signals might reside within them as described in the figure 2. Targets of these signals could be other RNAs, DNA or proteins. Long, single-stranded polyadenylated transcripts of satellite III are involved directly in the recruitment of splicing factors to nuclear stress granules (Metz *et al.*, 2004; Chiodi *et al.*, 2004). In addition, transcripts of st DNAs in the form of small interfering RNAs (siRNA) are thought to participate in the epigenetic process of chromatin remodulation and heterochromatin formation (Volpe *et al.*, 2002; Verdel *et al.*, 2004).

2.2.2 Simple Sequence Repeats (SSRs)

The SSRs can be found in both coding as well as non-coding genomes of prokaryotes and eukaryotes, and are present even in the small bacterial genomes (Toth *et al.*, 2000). SSRs also called as 'minisatellites' and 'microsatellites' are mutation-prone DNA tracts composed of tandem iterations of relatively short motifs than satellites. **Minisatellites** are moderately repetitive, tandemly repeated arrays of the moderately-sized (9 to 100 bp, but usually ~15 bp) repeats, generally

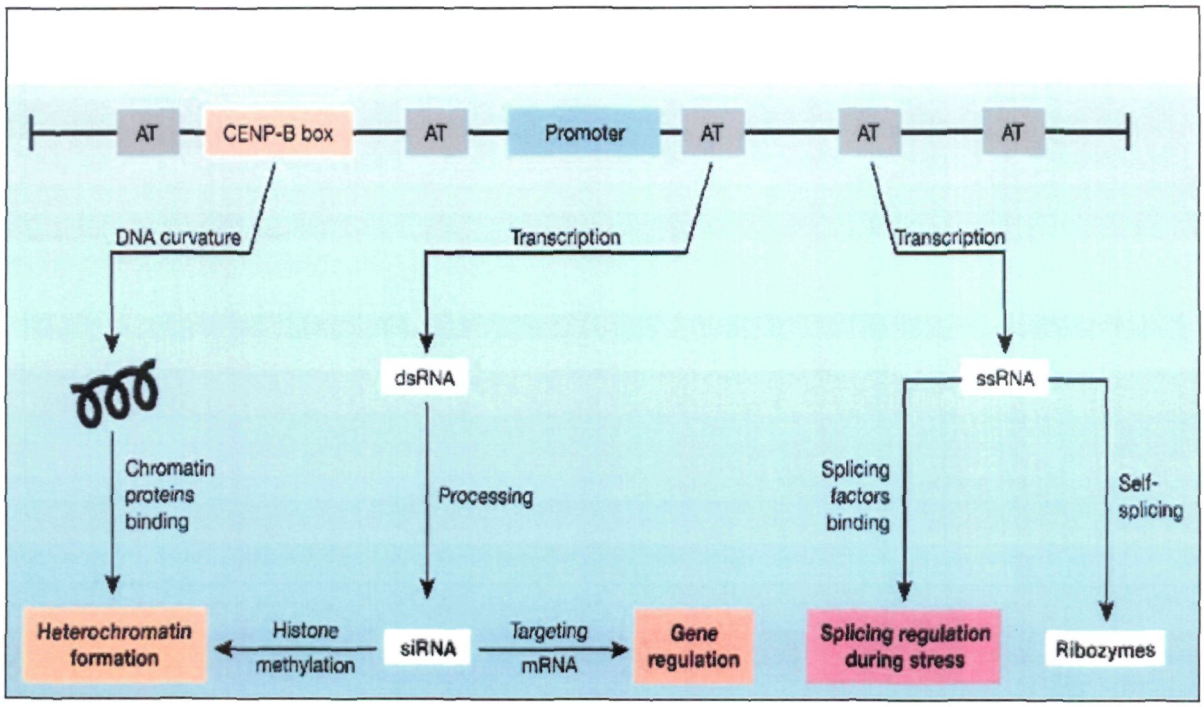


Figure 2. Schematic representation of a satellite monomer and the processes in which satellite DNAs are involved. Common functional elements such as periodically distributed AT tracts, centromere protein B (CENP-B) box and a promoter are indicated. A super-helical structure induced by DNA curvature, together with chromatin proteins such as CENP-B, could contribute to heterochromatin formation. Transcription of satellite DNAs proceeds either in the form of double-stranded RNA (dsRNA) or single-stranded RNA (ssRNA). dsRNA is processed into small interfering RNAs (siRNAs) that participate in heterochromatin formation and control expression of sequences tagged with complementary repeats. Single-stranded transcripts with hammerhead-like secondary structure have self-cleavage activity and function as ribozymes. Some ssRNAs, such as human satellite III stress-induced transcripts, specifically recruit splicing factors to nuclear stress granules and regulate splicing.

involving mean array lengths of 0.5 to 30 kb, whereas **Microsatellites** are moderately repetitive, and composed of arrays of short (2-6 bp) repeats. The human genome contains at least 100,000 microsatellite loci located in the euchromatin (Kashi and King, 2006). SSRs have been developed into one of the most popular classes of genetic markers owing to their high reproducibility, multi-allelic nature, co-dominant mode of inheritance, abundance and wide genome coverage (Schlotterer, 2004). Despite their ubiquitous occurrence, microsatellite density and distribution vary markedly across the genomes (Dieringer and Schlotterer, 2003). Because of the high mutability, SSRs are thought to play a significant role in genome evolution by maintaining and creating quantitative genetic variation. This genetic variation occurs primarily by slipped-strand mispairing and subsequent error(s) during DNA replication, repair or recombination (Levinson and Gutman, 1987), creating repeated tandem arrays called as variable number of tandem repeats (VNTRs).

2.2.2.1 Evolution of SSRs

The wide distribution of SSRs is a result of the dynamics and selective constraints of the human genome (Morgante *et al.*, 2002). The SSR abundance, various functions and effects (either putative or reliably established) are associated with their mutation rate. Although the mutation process seems to display distinct differences among species, repeat types, loci and alleles, age and sex (Brock *et al.*, 1999; Hancock, 1996; Ellegren, 2004; Schlötterer, 2004), the instability is predominantly manifested as changes in the number of SSR repeats. The repeat expansion/shrinkage processes also lead to the increase of biological complexity, which is considered to be the hallmark of biological evolution. Two mutational mechanisms can be invoked to explain such high rates of mutation. The first involves slippage during DNA replication called as 'DNA slippage' or 'slipped strand mispairing' (Tachida and Iizuka, 1992) and second involves recombination between DNA strands (Harding *et al.*, 1992). The efficiencies of the two mechanisms may putatively depend on the environmental conditions. Various factors are known to affect the rate of mutations at SSR loci including repeated motif, allele size, chromosome

position, GC content in flanking DNA, cell division (mitotic vs. meiotic), sex, and genotype (e.g. mutations at MMR genes). In a variety of widely diverged eukaryotes, including *S. cerevisiae, S. pombe, C. elegans*, Drosophila, plants, primates, and *Mus*, both coding and noncoding triplet SSRs are subjected to similar rates of repeat expansion (Metzgar *et al.*, 2000). SSR evolution in coding genes and regulatory regions should share mutational processes similar to those of the SSRs in untranscribed regions.

2.2.2.2 Characteristics of SSRs

The main characteristics of the simple repeats have been described below:

- *SSRs are extremely variable*

  Mutation size can vary from single base pairs at mononucleotide repeats up to multiples of much longer motifs in minisatellite repeats. SSR mutation rate is affected by motif length, motif sequence, number of repeats and purity of repetition (Armour *et al.*, 1999; Chambers and MacAvoy, 2000; Vergnaud and Denoeud, 2000; Ellegren, 2004). Point mutations can degrade repeat purity and stabilize an SSR; whereas active mutational slippage tends to eliminate imperfect repeats. Therefore, SSRs represent sites where selection can indirectly shape the site specific mutation rates at which new alleles arise.

- *Most SSRs are polymorphic*

  In the human genome for example, the proportion of AC repeats that are polymorphic is estimated to exceed 90% (Rockman and Wray, 2002). SSR polymorphism is known as the basis for DNA fingerprinting, lineage analysis and gene mapping. Normal variation in repeat number can be functionally significant. The number of repeats at particular SSR loci can influence several aspects of genetic function (Kashi and King, 2006); although small allelic differences in repeat

number commonly exert small quantitative phenotypic effects (many alleles can be effectively neutral).

- *They are ubiquitous and diverse*

  SSRs based on many different motifs are found in genomes of all species examined. They are abundant in various functional domains, both coding and non-coding. They occur within many open reading frames, but are even more frequent in non-coding regulatory regions (Rockman and Wray, 2002). Many genes are associated with more than one SSR; those containing at least one coding SSR often contain two or more (Karlin *et al.*, 2002; Kashi and King, 2006; Toth *et al.*, 2000) conducted a detailed analysis of SSRs in several eukaryotic taxa, from fungi to humans, and revealed highly taxon-specific patterns in the distribution of different repeat types (from mono- upto hexanucleotides) for different motifs in the coding and non-coding sequences, introns and intergenic regions. This specificity can partly be explained by interaction of mutation mechanisms and differential selection.

The accumulated empirical evidence seems to indicate that SSRs are more abundant and longer in vertebrates than in invertebrates. Among vertebrates, longer SSR tracts are observed in cold-blooded species (Chambers and MacAvoy, 2000). It is interesting that among the texa compared by Toth *et al.*, 2000, maximum abundance of SSRs was displayed by rodents and the minimum by *C. elegans.*

2.2.2.3 Distribution of SSRs in the coding and non-coding genomes

The frequency of distribution of the SSRs with different motifs varies by functional domain, with triplet motifs much more common within coding regions (Toth *et al.*, 2000; Katti *et al.*, 2001; Morgante *et al.*, 2002; Ellegren, 2004). All types of other SSRs (from mono to hexanucleotide repeats) were found in excess in non-coding regions across seven eukaryotic clades: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*,

Schizosaccharomyces pombe, *Mus musculus*, *Drosophila,* plants, and primates (Metzgar *et al.*, 2000). Morgante *et al.*, 2002 reported that all SSR types except triplets and hexanucleotides are significantly less frequent in the 25,762 predicted protein-coding sequences compared with the non-coding fraction in six plant species including *Arabidopsis*, rice, soybean, maize, and wheat (*Triticum aestivum*). In the genome of Japanese pufferfish, *Fugu rubripes*, only 11.6% of 6042 SSRs were detected in the protein-coding regions (Edwards *et al.*, 1998). This is attributable to negative selection against frameshift mutations in coding regions (Metzgar *et al.*, 2000). Previously, a similar distribution pattern was found for triplet SSRs in coding and non-coding genomes of fungi, protists, prokaryotes, viruses, and humans (Field and Wills, 1998; Wren *et al.*, 2000). However, the disease-associated triplet repeats are mostly found in coding regions of the human genome (Nadir *et al.*, 1996). Likewise, Morgante *et al.,* 2002 recently found that triplet SSRs doubled in frequency in the coding region of the above-mentioned six plant species, as a result of mutation pressure and possibly positive selection for specific single amino acid stretches. In contrast to the triplet SSRs, di-and tetranucleotide o are much less frequent in coding than in the noncoding regions. For example, dinucleotide repeats are about 20 times less frequent in the expressed sequences than in random genomic clones of Norway spruce, *Picea abies* (Scotti *et al.*, 2000). In eight prokaryotes and yeast, long mono- and di-tracts are almost exclusively distributed in non-translated regions (Field and Wills, 1998). Different species have different motif distributions frequency. The majority of SSRs (48-67%) found in many species are dinucleotides (Wang *et al.*, 1994; Schug *et al.*, 1998) for e.g. in human, *Caenorhabditis elegans* and *Arabidopsis thaliana* genomes are ACn, AGn and ATn, respectively (Toth *et al.*, 2000; Katti *et al.*, 2001). In contrast, the primate mononucleotides (mainly, poly(A/T) tracts) are the most copious classes of SSRs (Toth *et al.*, 2000; Wren *et al.*, 2000).

The differences between coding and non-coding SSR frequencies arise from specific selection against frame-shift mutations in coding regions resulting from length changes in nontriplet repeats (Liu *et al.*, 1999; Dokholyan *et al.*, 2000). Nevertheless, 14% of all proteins contain

repeated sequences, with a three times higher abundance in eukaryotes compared to prokaryotes (Marcotte *et al.*, 1999). Prokaryotic and eukaryotic repeat families are clustered to non-homologous proteins. This may indicate that repeated sequences have emerged after these two kingdoms split. The eukaryotes incorporating more repeats may have an evolutionary advantage of faster adaptation to new environments (Marcotte *et al.*, 1999; Kashi *et al.*, 1997; King and Soller, 1999; Wren *et al.*, 2000).

2.2.2.4 Putative functions and effects of SSRs in genome

2.2.2.4.1 In the Non-coding Genome

Although SSRs were usually considered just as evolutionary neutral DNA markers, the functional significance of a substantial part of SSRs has been proven by critical tests in various biological phenomena, as shown in the figure 3.

2.2.2.4.1.1. Chromatin organization

- *Chromosomal organization*

  Some aspects of SSR distribution point to their possible role in taxon-specific chromosome structure. For instance, SSR hybridization signals were found in related chromosome positions independently of the motif used, and showed remarkably similar distribution patterns in wheat and rye, suggesting a special role of SSRs in chromosome organization (Cuadrado and Schwarzacher, 1998).

- *DNA structure and packaging*

  SSRs are capable of forming a wide variety of unusual DNA structures with simple and complex loop-folding patterns. For example, the hairpin formed by the fragile X repeat (CCG)*n*, and the bipartite triplex formed by (GAA)*n* /(TTC)*n*, show simple loop folding. Such triplex structures may have important regulatory effects on gene expression (Fabregat *et al.*, 2001). The human centromeric repeat (AATGG)*n* can form a double-folded hairpin DNA structure (Catasti *et*

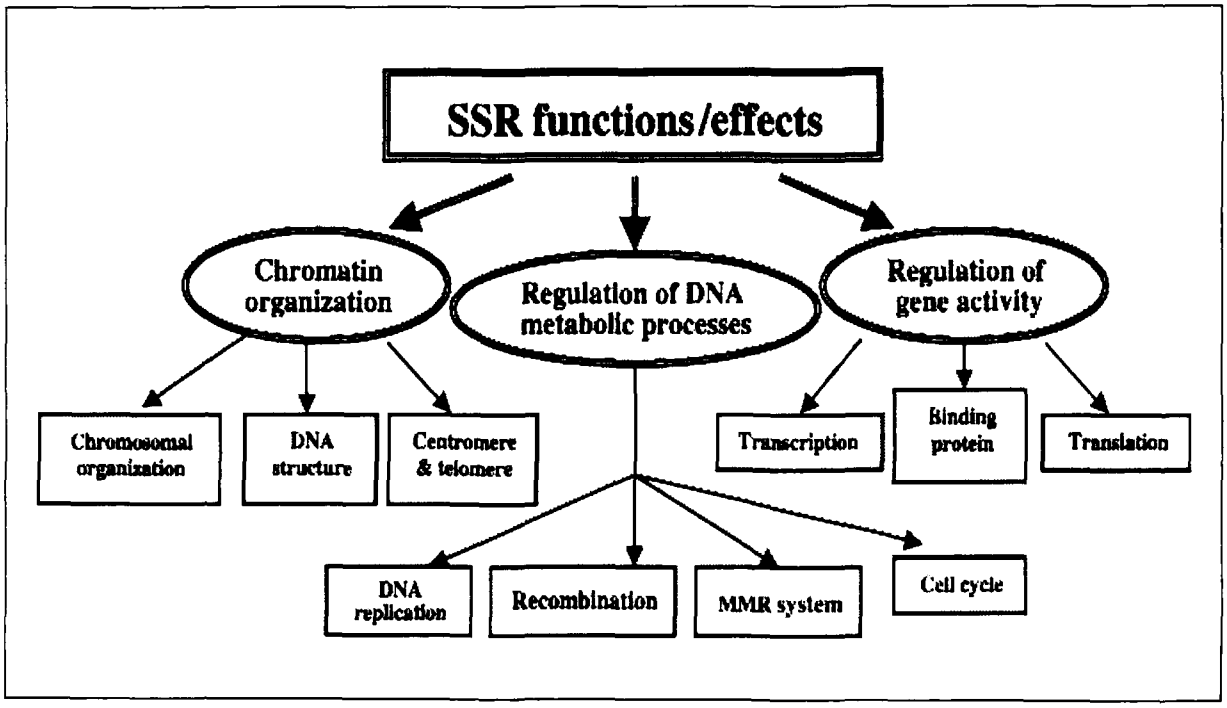


Figure 3. The functions and effects of the Simple Sequence Repeats (SSRs) in the non-coding regions of the genome. Major functions included chromatin organization, regulation of DNA metabolic processes and gene activity regulation. The sub-branches of these functions has been demonstrated in the figure.

al., 1999). The formation of such stable structures offers a mechanism of unwinding which is advantageous during transcription, and provides unique protein recognition motifs (Catasti *et al.*, 1999). In many species, dimeric SSR relative abundance may also reflect duplex curvature, supercoiling, and other higher-order DNA structural features (Baldi and Basnee, 2000).

- *Centromere and telomere*

  In many species, the chromosomal centromeric region is composed of numerous tandem repeats, which affect the centromere organization. Long SSRs with mono-, di-, tri- and tetranucleotide motifs are highly clustered in the centromeric regions of tomato (Areshchenkova and Ganal, 1999), *Arabidopsis* (Brandes *et al.*, 1997), and *Beta vulgaris* (Schmidt and Heslop-Harrison, 1996). In *Drosophila* minichromosomes, the centromere flanking DNA predominantly contains highly repetitive sequences, and the number of repeats required for normal transmission differs among cell division types and between the sexes (Murphy and Karpen, 1995). The centromere flanking repeated DNA may assist in sister chromatid cohesion and kinetochore formation.

2.2.2.4.1.2. Regulation of DNA metabolic processes

- *Recombination*

  Numerous SSRs have been proposed as hot spots for recombination (Jeffreys *et al.*, 1998; Templeton *et al.*, 2000). Support for this idea was provided by experiments with simian virus 40 (Wahls and Moore, 1990a), yeast (Treco and Arnheim, 1986), human (Aharoni *et al.*, 1993; Majewski and Ott 2000; Templeton *et al.*, 2000), and mammalian cells (Wahls and Moore, 1990b). Dinucleotide repeat sequences are preferential sites for recombination because of their high affinity for recombination enzymes (Biet *et al.*, 1999). Some SSRs may influence recombination directly by their effects on DNA structure. It has been proposed that GT, CA, CT, GA, GC, or AT repeat-binding

proteins could participate in recombination processes by inducing Z-conformation or other alternative secondary DNA structures (Biet *et al.*, 1999). The repeat number is also known to influence the recombination process (Dutreix, 1997).

- *DNA replication and cell cycle*

  Few studies suggest the effects of SSRs on DNA replication (Field and Wills, 1996). In rat cells, DNA amplification is arrested within a specific fragment, consisting of a $d(GA)_{27}$,d(TC)27 tract. This tract is found at the end of an amplicon, and in conjunction with the inverted repeat, may serve as an arrest site for DNA replication *in vivo*. SSR can also affect enzymes controlling cell cycles. For instance, the human *CHK1* gene has a role in controlling cell cycle progression, and its coding region contains an (A)9 tract (Codegoni *et al.*, 1999) that is a potential site of mutations in tumors with SSR instability (Bertoni *et al.*, 1999). Alterations in the *CHK1* gene in human colon and endometrial cancers were associated with the presence of a high degree of poly(A) tract instability leading to the truncated protein. Alterations of the *CHK1* gene could represent an alternative way for cancer cells to escape from cell cycle control (Bertoni *et al.*, 1999). Some genes controlling the cell cycle, such as *hMSH3*, *hMSH6*, *BAX*, *IGFIIR*, *TGFbetaIIR*, *E2F4* and *BRCA2*, carry short repeated sequences, important in cell fidelity and growth control. SSR instability affects these genes by both insertions and deletions of repeat units (Johannsdottir *et al.*, 2000).

- *SSRs in the eukaryotic DNA as modulators of evolutionary mutation rate*

  DNA mismatch repeair (MMR) proteins correct replication errors and actively inhibit recombination between diverged sequences (Kolodner and Marsischky, 1999), thus controlling mutation rates and evolutionary adaptations. It is found that the constellation of (A)*n* SSRs in the coding regions of the minor MMR genes (*MSH3*, *MSH6*, *PMS2* and *MLH3*) is a general feature among different eukaryotes and

prokaryotes. SSRs are exceptionally vulnerable to spontaneous insertion or deletion mutations, and the nontriplet SSRs, when located in coding sequences, are expected to introduce frameshift loss of function mutations at high frequency (Moxon *et al.*, 1994). Chang *et al.*, 2001 hypothesized that the exceptional density of SSRs in the minor MMR genes represents a genetic switch that allows the adaptive mutation rate to be modulated over evolutionary time.

2.2.2.4.1.3. Regulation of gene activity

- *SSRs and transcription*

Evidences show that SSRs located in promoter regions may affect gene activity. The (TC)*n* tract in promoter regions serve as a transcriptional element for heat-shock protein gene *hsp26* in *Drosophila* (Sandaltzopoulos *et al.*, 1995). Deletions of various di, tri and tetra-SSR tracts markedly change transcriptional activity (Hoffman *et al.*, 1990). SSRs in intronic regions can also affect gene transcription (Meloni *et al.*, 1998). Transcription activity of the epidermal growth factor receptor gene declines with increasing numbers of (CA)n repeats (Gebhardt *et al.*, 1999; 2000). It is noteworthy that triplet SSRs may be preferentially located in regulatory genes related to transcription and signal transduction, and remains under-represented in genes for structural proteins (Young *et al.*, 2000), suggesting effects of an SSR on gene transcription.

- *Implication of repeat number variation on gene expression*

SSR numbers appears to be a key factor for the regulating the mechanism and level of gene expression. Some genes can only be expressed at a specific repeat number of SSRs (Liu *et al.*, 2000). Others can be expressed within a narrow range of SSR numbers, and out of this range, gene activity would be turned off. In a CAT reporter system carrying an androgen response element with human CAG repeats in the presence of dihydrotestosterone, expansion mutations showed a progressive decrease in transcriptional transactivation with

increasing CAG repeat length (Chamberlain *et al.*, 1994). In contrast, some genes' transcriptional levels increase with the SSR numbers (Okladnova *et al.*, 1998). The experiments with various organisms, indicates the importance of SSR numbers for SSR-related regulation of gene expression.

- *Protein binding and Translation*

  Some SSRs, found in upstream activation sequences, serve as binding sites for a variety of regulatory proteins (Csink and Henikoff, 1998). For example, single-stranded poly(GA)- and poly(GT)-binding proteins have been identified in human fibroblasts (Aharoni *et al.*, 1993). SSR repeat number may also affect protein binding (Winter and Varshavsky, 1989).

  Many studies showed the effect of SSRs on gene translation. For instance, a moderate size expansion of CGG tract can markedly reduce translation of *CAT* gene (Sandberg and Schalling, 1997). The distribution of AGCT tetranucleotides in the *E. coli* and *Bacillus subtilis* genomes predicts translational frameshift and ribosomal hopping in several genes (Henaut *et al.*, 1998). Also, *CAT* reporter gene has proven a strong inhibitory effect of AGG triplet repeats on translation in *E. coli* (Ivanov *et al.*, 1992).

2.2.2.4.2 In the Coding Genome

Some of the SSRs are retained within transcripts from their `birth' in the nucleus to their `death' in cytoplasm. Few reports are there only to speculate the potential functions of these repeats in coding genome being involved in many steps of regulation which can be broadly categorized as:

- *RNA shape and regulation*

  Many cellular transcripts form a multitude of various hairpins, and few of them are composed of triplet repeats. Jasinska and group have investigated 20 transcripts composed of triplet repeat motifs containing 17 iterations, were found to form hairpin structures. Among them, CNG

repeats were most abundant in transcripts (Toth *et al.,* 2000). The characteristic feature of CNG repeat hairpins is that their structure rigidity increases with repeat length (Cohen *et al.*, 2004). However, the repeated sequences may be hidden or exposed in these transcript structures for interactions with RNA binding proteins for e.g. hairpin folds of RNA transcribed from CTG repeats in 3' UTR of myotonic dystrophy protein kinase gene bind to and activate the protein kinase (Tian *et al.*, 2000). A variable TG repeat in the cystic fibrosis transmembrane conductance regulator gene (CFTR) alters the efficiency of exon splicing (Hefferon, 2004).

- *SSR Variation, Gene Expression and Pathogenesis in Prokaryotes*

  The presence of SSRs in prokaryotes is rare, but the few reports are related to pathogenic organisms and the variation in their repeat numbers can cause phenotypic changes (van Belkum *et al.*, 1998). These SSR motifs were reminiscent of the presence of repetitive elements consisting of uptake signal sequences, intergenic dyad sequences, and multiple tetranucleotide iteration (Karlin *et al.*, 1997). *Haemophilus influenzae* (Hi), an obligate upper respiratory tract pathogen, uses phase variation (PV) to adapt to host environment changes. Switching occurs by slippage of SSR repeats within genes coding for virulence molecules (Hood *et al.*, 1996). Moreover, variation in the opacity surface proteins (Opa) in the coding repeat sequence causes the shifting of the translational reading frame (Murphy *et al.* 1989). SSR variations enable bacteria to respond to diverse environmental factors, and many of them are clearly related to bacterial pathogenesis and virulence.

- *Effects of SSR on Gene Expression or Gene Silencing*

  Gene expression is crucial for maintenance of differentiated cell types in multicellular organisms, whereas aberrant silencing can lead to disease. SSR elements in the 5'-UTRs are required for some gene expression. For instance, the human calmodulin-1 gene (hCALM1)

contains a stable (CAG)7 repeat in its 5'-UTR (Toutenhoofd *et al.*, 1998). Experiments have demonstrated that deleting this repeat decreases the gene expression by 45%, whereas repeat expansions to 20 and 45 repeats, or the insertion of a scrambled (C, A, G)7 sequence did not alter gene expression (Toutenhoofd *et al.*, 1998). SSRs in 3' UTR and introns of few genes has been reported to silence the gene function e.g. short GAA or CTG repeat expansions in 3'-UTR of DM1 and the intron of FRDA gene mediate heterochromatin-protein-1–sensitive variegated gene silencing (Saveliev *et al.*, 2003).

- *Transcriptional slippage and Translational Control*

  Transcription of a CAG/CTG triplet repeats in 3'-UTR of a URA3 reporter gene in yeast leads to transcription of mRNA several kilobases longer than the expected size. These large mRNA molecules are formed by transcription slippage (Fabre *et al.*, 2002), a phenomenon reported to be usually stimulated by short mononucleotide and dinucleotide repeats (Davis *et al.*, 1997). CAG/CTG repeats form secondary structures in vitro (Gacy *et al.*, 1995) and stable *in vivo* secondary structures formed are important for transcription slippage.

  SSRs in 5'-UTRs serve as protein binding sites, thereby regulating gene translation, protein component and function. For instance, the transcription factor CCAAT/enhancer binding protein b (C/EBPb) plays a significant role in the regulation of hepatocyte growth and differentiation. Different RNA binding domains of a CUG repeat binding protein (CUGBP1) bind to both the CUG and CCG repeats on this gene. The binding of CUGBP1 to the 5' region of C/EBPb mRNA results in generation of low molecular weight C/EBPb isoforms. It is possible that this interaction may stabilize a structure that favors translational initiation at downstream AUG codons (Timchenko *et al.*, 1999).

2.2.2.5 SSRs and diseases linkage

Repeat variation within genes should be very critical for normal gene activity because the SSR expansion or contraction directly affects

the corresponding gene products and even causes phenotypic changes. In eukaryotes, repeat instability has been studied as an important and unique form of mutation that is linked to more than 40 neurological, neurodegenerative and neuromuscular disorders. Human repeat expansion diseases are predominantly caused by instability and expansion of triplet motifs within or near genes (Cummings and Zoghbi, 2000). Of the 40 SSR related disorders, 16 are known to be caused by expansions of trinucleotide repeats (CUG)n, (CGG)n, (CCG)n, (GAA)n and (CAG)n in single genes and the largest class of these diseases results from CAG repeats, translated into extended (Gln)n tracts within the corresponding proteins. However, the dominant repeat-associated disorders include Huntington disease (HD), dentatorubro-pallidoluysian atrophy (DRPLA), myotonic dystrophy types 1 and 2 (DM1 and DM2), fragile X syndrome (FRAXA), spinal bulbar muscular atrophy (SBMA), Friedreich's ataxia (FRDA), a series of spinocerebellar ataxias (SCA1–3, 6, 7, 8, 10, 12 and 17), epilepsy, progressive myoclonic 1 (EPM1), insulin (INS) and facioscapulohumeral muscular dystrophy 1A (FSHMD1A).

2.3 Minisatellite/Microsatellite Associated Sequence Amplification (MASA) approach

MASA, an advanced form of Random Amplified Polymorphic DNA fragment (RAPD) or Arbitrarily Primed Polymerase Chain Reaction (AP-PCR), employs the random primers to amplify the random segments of genomic DNA or transcripts to reveal identification of repeat tagged genome and polymorphisms present therein. However, RAPD uses the primer sequences which are confined to a single or limited number of genomes, whereas MASA utilizes the evolutionary conserved sequences represented by minisatellites or microsatellites (Williams *et al.*, 1990). These sequences have been implicated with high rate of recombinational activities leading to sequence polymorphism in the genome. MASA has been used to assess species-specific band profile in *Rhinoceros unicornis* (Ali *et al.*, 1999; Kapur *et al.*, 2003) and male specific band pattern in the humans (Bashamboo and Ali, 2001) using these hypervariable evolutionary repeats. Thus, amplicons generated by MASA with

evolutionarily conserved primer(s) or sequences shared by many species are particularly useful for clad identification in controversial systematics, comparative genome analysis, and for establishing the phylogenetic status of an organism in addition to wildlife conservation biology and forensic medicine.

The evolutionary conserved repeats used in the present study were the consensus of 33.15 repeat loci, and simple repeats of GACA and GATA sequences. The 16 bp nucleotide hypervariable repeat of 33.15 sequence was originated from human myoglobin gene (Chromosomal position: 7q35-36) (Jeffereys *et al.*, 1985) whereas quadruplet (GACA)n and (GATA)n repeats first observed in the banded krait minor (Bkm) satellite DNA of female snakes, *Bungarus fasciatus* (Singh *et al.*, 1980) and *Elaphe radiata* (Epplen *et al.*, 1982). The 33.15 repeat reported from a number of species has been extensively used for genome polymorphism due to its hypervariable nature (Ali and Wallace, 1988; Jeffereys *et al.*, 1985). This repeat has also been shown to associate with heterochromatic sequences of the human Y chromosome (Bashamboo and Ali, 2001). Other two simple repeats of GACA/GATA were found to be arranged in a sex-specific manner in mouse, fish, reptile and avian genomes but in a non-sex manner specifically in other eukaryotes including the humans (Singh *et al.*, 1980; Epplen *et al.* 1982; Hobza *et al.*, 2006). These repeats also showed sex- and tissue- specific expression in higher eukaryotes supporting their crucial role in sex-differentiation (Singh and Jones, 1982; Subramanian *et al.*, 2003). Though the amount and organization of these vary considerably, but they seem to be present in almost all the eukaryotes studied. All of these sequences have been used to for DNA fingerprinting or directly as hybridization probes but have not been used as primers to amplify the genes or transcripts in context of genome identification.

Recent profound analysis of the animal genomes has inspired interests in methods for molecular characterization of DNA and mRNA transcripts. Several complementary avenues for the study of mammalian genomes are currently being considered. Some strategies are directed towards creating high-resolution linkage maps of representative genomes,

whereas others emphasize the development of coding maps of cloned sequences. We have focused on one aspect of these strategies for genome analysis, i.e. the assessment of organizational status of the buffalo genome and transcriptome using these simple sequence repeats. Present study is an attempt to generate information on the distribution pattern of tandem repeats among transcripts originating from different tissues and spermatozoa and assess their possible role in the gene regulation, besides characterization of these genes with respect to their organization and expression in different tissues in buffalo *Bubalus bubalis*.

MATERIALS AND METHODS

3. MATERIALS AND METHODS

3.1 Sample collection and genomic DNA isolation

Peripheral blood and tissue samples from both the sexes of water buffalo (*Bubalis bubalis)* were collected with the help of veterinarians on duty from local slaughter house, Delhi following strictly the guidelines of Institutional Ethical and Bio-safety Committee. Fresh ejaculates from the buffaloes were collected from the local dairy farm. Genomic DNA was isolated from blood and semen samples using phenol: chloroform: Isoamyl-alcohol extraction method (Srivastava *et al.*, 2006a; 2006b). Prior to DNA isolation, semen samples were washed thrice with sperm wash buffer (0.15 mM NaCl, 10mM EDTA; pH=8.0). For cross hybridization studies, DNA was also extracted from peripheral blood of cattle *Bos indicus*, sheep *Ovis aries*, goat *Capra hircus*, human *Homo sapiens,* Pigeon *Columba livia*, pig *Sus scrofa*, Baboon *Papio hamadryas*, Bonnet monkey *Macaca radiata*, Langur *Presbytis entellus*, Rhesus monkey *Macaca mulatta*, Lion *Panthera leo,* Tiger *Tigris tigris.* Lion and Tiger blood samples were procured with due approval of the competent authorities of the States and Union Government of India. Tissues were homogenized before DNA isolation procedure. Blood samples and powdered tissue were mixed in the lysis buffer and kept on ice for 15 minutes with intermittent mixing. The samples were pelleted and the pellet was resuspended in 5 ml nuclease buffer with 0.2% SDS and 0.02 mg/ml proteinase K and incubated at 37^0C for 4-16 hr. It was then extracted once with Tris-saturated phenol, once with phenol: chloroform: isoamyl alcohol (25:24:1) and twice with chloroform: isoamyl alcohol (24:1). The DNA from the supernatant was precipitated by $1/10^{th}$ volume of sodium acetate and twice volume of distilled ethanol followed by pelleting, and washing the pellet twice with 70% ethanol. The concentration of the DNA was estimated by spectrophotometer using the formula:

Concentration of DNA = Optical Density (OD) $_{260\ nm}$ x Dilution Factor (DF) x 0.04 mg/ml

Where 1 OD = 0.04 mg/ml (for ss DNA and RNA)

3.2 Isolation of total RNA and cDNA synthesis

For RNA isolation from blood, first Peripheral blood mononuclear cells (PBMCs) were isolated using Histopaque gradient method as per suppliers' specifications (Sigma Aldrich). Briefly, around 3 ml of the blood was layered gently on the equal volume of the histopaque-1077 at room temperature followed by centrifuge at 600 g for 30 minutes. The PBMCs appearing in the form of a white translucent ring were taken from between the lower RBCs and upper serum layers. Total RNA was isolated from PBMCs and different tissues (testis, ovary, spleen, liver, lung, kidney, brain and heart) using Tri-X Reagent (Molecular Research Center, Cincinnati, OH) following supplier's specifications. The tissues were homogenized in TRI-X reagent, kept at room temperature for 5 minutes and the aqueous layer containing RNA was extracted using chloroform separation method (Srivastava *et al.*, 2006a; 2006b). Semen samples were subjected to percoll gradient method to select only motile sperms (Morales *et al.*, 1991). Sperms were washed using sperm wash solution (0.15mM NaCl, 10mM EDTA pH 8.0) twice following by RNA isolation as described previously (Miller *et al.*, 1994). The RNA was then treated with RNase-free DNase-1 (10 U in 50 mM Tris-HCl, 10 mM $MgCl_2$, pH 7.5) and then re-extracted. Total RNA samples were checked on 1% agarose gel in 1X MOPS buffer (10X MOPS contains 0.2 M 3-N-(morpholino) propanesulfonic acid; 0.05 M sodium acetate and 0.01 M EDTA, pH adjusted to 7.0). The cDNA synthesis was conducted using a commercially available kit (GIBCO-BRL, Gaithersburg, MD) and confirmed by PCR amplification using a set of bubaline derived ß-actin (forward 5' GTGG GCCGCTCTAGACACCA 3' and reverse 5' CGGTTGGCCTTAGGGTCG GGGGG 3') primers. Quantification of RNA and cDNA was done by UV spectrophotometer at 260 nm. Final RNA preparations were tested for residual DNA contamination by PCR using primers against ß-actin and isolated RNA as template following standard procedures.

3.3 Minisatellite/Microsatellite Associated Sequence Amplification (MASA)

To conduct MASA reactions, 6 sets of oligos based on the GACA and GATA repeats, and a 16-nucleotide long oligo based on the consensus of the 33.15 repeat loci (Table 1), respectively, were designed, and purchased from Microsynth GmbH (Balgach, Switzerland). MASA reactions were individually performed for each oligo using cDNA samples as template from different tissues and spermatozoa in a 25-50 µl reaction volume each containing 2-4 µl of cDNA (corresponding to about 2-4 ng purified mRNA), 20-40 pmole of primer for each repeat (Srivastava *et al.*, 2006a; 2006b). Briefly, the template was first denatured at 95°C for 5 min followed by 35 cycles each with a subsequent denaturation at 95°C for 1 min, annealing at corresponding temperatures and extension for 1-2 min each at 72°C, followed by a final extension at 72°C for 10 min. The sequence and annealing temperature for each oligo primer has been given in the table 1. The resultant amplicons were resolved on 20 cm long 2% (w/v) agarose gel using 0.5X TBE buffer.

3.4 Cloning, Sequencing and Characterization of the MASA uncovered amplicons

From the MASA reactions, 148 amplicons were uncovered using consensus sequence of 33.15 repeat, 332 amplicons with GACA repeat motif and 136 amplicons with GATA motif. These amplicons resolved on the agarose gel were sliced; DNA was eluted (Qiagen Gel Extraction kit, Germany) and processed independently for ligation into pGEMT-easy vector (Promega, USA) using standard protocols. The competent cells were prepared using $CaCl_2$ treatment method (Sambrook *et al.*, 1989).The ligation product was transformed into the DH5α competent cells by heat shock method. The recombinant clones were screened by blue/white selection. The positive clones were authenticated by restriction digestion, Slot blot and Southern hybridization using labeled buffalo genomic DNA following standard methods (Sambrook *et al.*, 1989). Recombinant plasmids purified using alkali-lysis methods were subjected to sequencing. The resultant recombinant clones were sequenced and the sequences

Table 1: List of oligos based on the repeat sequence motifs of GACA and GATA, and the consensus of 33.15 repeat loci for the identification of their tagged transcripts

S.N.	Oligo ID	Sequence (5'-3')	Annealing Temp. (in ^{O}C)
1.	OAT15.2	5' ACAGACAGACAGACA 3'	55.5
2.	OAT18.2	5' ACAGACAGACAGACAGACA 3'	56.8
3.	OAT24.2	5' ACAGACAGACAGACAGACAGACAGACA 3'	60
4.	OAT15.1	5' ATAGATAGATAGATA 3'	53.6
5.	OAT18.1	5' ATAGATAGATAGATAGATA 3'	54.9
6.	OAT24.1	5' ATAGATAGATAGATAGATAGATAGATA 3'	59.5
7.	OAT33.15	5' CACCTCTCCACCTGCC 3'	59.0

were deposited in the GenBank. The accession numbers were obtained for each of the submitted sequences and have been described in detail in the results section.

3.5 Cloning and characterization of buffalo *c-kit* and *Smoc*-1 genes

Four sets of overlapping PCR primers encompassing different regions (Table 2) specific to buffalo *c-kit* gene were designed based on mRNA sequence of *Bos taurus* and *Bos primigenius* (Accession numbers AF263827 and D16680, respectively) using primer3 output (http://frodo.wi.mit.edu/cgi-bin/primer3 /primer3www.cgi). Details of the primer sequences, Tm and corresponding product sizes are given in the table 2. PCR reactions were conducted using Vent or pfu+taq Polymerases (NEB, USA) involving denaturation at 95^0C for 5 min, followed by 40 cycles of 95^0C for 1 min, Tm ^{0}C for 2 min, 72^0C for 2 min and final extension at 72^0C for 10 min. The PCR products were tagged with dTTP at the 3' ends using *Taq polymerase,* and cloned into *pGEM*T-Easy vector (Promega, USA). The recombinant plasmids were confirmed by various approaches as discussed in the section 3.4.

Full length *Smoc*-1 was isolated using four sets of primers designed from of 5' and 3' regions of the human and cattle *Smoc*-1 sequences (GenBank Accession nos. AJ249900 and XM_612029) respectively using primer3 output (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3www.cgi). Details of the primer sequences, Tm and corresponding sizes of amplicons are given in the table 3. End point PCR was conducted to amplify 3'UTR using cattle derived primers (JS275-JS997). The 5' UTR and polyadenylation signal at 3'UTR were identified using 5' & 3' RACE kits (Invitrogen, USA) and *Smoc*-1 specific primer J5UTR & JS977/JS997, respectively. All the PCR reactions were conducted using Vent Polymerase (NEB, USA) following standard protocols. The amplicons were analyzed by the gel electrophoresis followed by the cloning and sequencing. The sequences so obtained were assembled into full length *Smoc*-1 (FSmoc-1). To ascertain the possible insertion of 12^{th} intron in 3'UTR, the primers JS990-SA991 & JS996-997 were used on buffalo

Table 2: Details of Primers used for the isolation of the full length CDS of *c-kit* and expressional studies using Real Time PCR in Water buffalo *Bubalus bubalis*

Primers	Length	Nucleotide sequence (5'-3')	Annealing temperature (^{0}C)	Amplicon size (In bp)
Primers used for amplification of *c-kit* gene				
CK148	20	5' AACGATGTGGGCAAGAGTTC 3'	63	1437
CK149	20	5' GACATCTTCGTGGACAAGCA 3'		
CK174	18	5' AATGGGACGGTGGAGTGC 3'	62	1480
CK175	18	5' GGAGACCCCCAGATGCAG 3'		
CK176	18	5' CCGGAACGTGGAACAGAG 3'	65	1498
CK177	22	5' TGTTTGAATGTGCTGTCATCA A 3'		
CK150	20	5' ATAGCTGGCATCAGGGTGAC 3'	64	679
CK151	19	5' CCCATTGTGTTTGAATG3'T		
Primers used for assessment of copy number calculation and relative expression				
CRT1	19	5' CCAAGGCAGGCATCACAAT 3'	60	70
CRT2	19	5' TCGCTGAGCAGTGCAGACA 3'		

The size of oligos, their annealing temperature and corresponding product size of the respective amplicons have also been given.

Table 3: Details of Primers used for amplification of the full length *Smoc*-1 CDS, Its relative expression and copy number calculation#

S.N.	Clone ID	Primer code	Sequence(5'-3')	Mers	Annealing temp. (In ^{O}C)	Product size
For full length CDS amplification						
1.	Clone II	JS275	5' ATGACTGTGTCCCCTGACCGCAGC 3'	24	65	1414
2.		JS276	5' TCTGATGATCCATCCTGCCTCCCTGGT 3'	27	67	
3.	Clone III	JS985	5' TGTGACCTGAACAAGGACAAGGT 3'	23	62	1009
4.		JS986	5' TTCTGTGCTTCTACCTCCACTGC 3'	23	62	
5.	Clone IV	JS990	5' AACTTTCTTCACGGAGGTGCTTC 3'	23	61	1030
6.		JS991	5' AATTGCCCTAAGGATTTGCGTTA 3'	23	61	
7.	Clone V	JS996	5' AAATGTAATATCTGAGCAGTGGAGGT 3'	26	60	1000
8.		JS997	5' TATAAAACAAGCTACAAACGGTCTCC 3'	26	60	
9.	Clone VII	JS977	5' GTTTCACTGACTACTGTGACCTGAAC 3'	26	65	506
10.	Clone VIII	JS997	5' TATAAAACAAGCTACAAACGGTCTCC 3'	26	60	1018
11.	Clone VI	J5UTR	5' CAGCAGTACCCGGTGTAAGTATG 3'	23	63	649
For copy number calculation and relative expression						
10.	Variants-01+-02	JSR1015	F 5' CGCGTGGTGCACTGGTATT 3'	19	60	72
11.		JSR1016	R 5' CTTCATCTCGCGCTTGTTGA 3'	19	60	
12.		JSR1017	F 5' ATCAACAAGCGCGAGATGAAG 3'	21	60	74
13.		JSR1018	R 5' CGCCGGGCACATTTCTT 3'	17	60	
14.	Variant-01	JSR1033	F 5' AGGGAGGGTGGGCAGTTTT 3'	19	60	74
15.		JSR1034	R 5' TGCGGCCAGATTTCCAAA 3'	18	60	

The size of oligos, their annealing temperature and corresponding product size of the respective amplicons have also been given.

genomic DNA as template to conduct end point PCR. The PCR products were tagged with dTTP and cloned into *pGEM*T-Easy vector (Promega, USA) followed by sequencing. Finally, derived full length transcripts were submitted in the GenBank (Accession numbers: DQ159955 and EF446167).

3.6 Homology status, phylogenetic delineation, domain organization, secondary structure prediction and other in-silico analyses

The individual raw sequences were edited using Gene Runner software. Database search was conducted to determine homology of these sequences independently with other entries in the GenBank using default server (http://www.ncbi.nml.nih.gov.cgibin/BLAST/nph-newblast).

Restriction mapping of the recombinant clones was also done using NEB-cutter (www.tools.neb.com/NEBcutter2) and based on this information, *Rsa*I, *Hin*fI and *Hae*III enzymes (NEB, USA) were used to look for the absence/presence of any inter-clonal variation. The sequences of the two clones each from every single amplicon were independently subjected to ClustalW alignment (www.ebi.ac.uk/clustalw) to ascertain inter-clonal sequence variation. The secondary structure of predicted protein was ascertained using Phyre software (http://www.sbg.bio.ic.ac.uk/phyre). The calcium binding affinity, and N-glycosylation and O-glycosylation sites in the Smoc-1 were predicted using different bioinformatics tools. Based on the homology, the phylogenetic tree (http://sbg.bio.ic.ac.uk/phyre) was constructed using *c-kit* and *Smoc*-1 sequences from buffalo and other species. Only those species showing significant homologies (80-100%) were taken into consideration.

3.7 RNA slot blot analysis and Northern blotting

For RNA slot blot analysis, approximately 2 µg of total RNA from different tissues of buffalo was slot blotted onto a nylon membrane (Minifold Apparatus, Schleicher & Schuell, Germany) and UV fixed. For positive control, 5 ng of recombinant plasmid, each, was included in the blot(s). For Northern blot analyses, 5-10 µg total RNA was denatured on

65^0C for 10 minutes and separated on 1% agarose gel containing 4% formaldehyde and transferred to nylon membrane (Amersham Biosciences) using capillary transfer method. Individual probes for each fragment were labeled with [^{32}P] α-dCTP using random priming kit (Rediprime^{TM} II, Amersham Pharmacia biotech, USA). The pre-hybridization was carried out for 4-6 hours at hybridization temperature of the individual probes using 1ml/cm2 of hybridization fluid [5X Denhardts solution (0.5 g Fiicoll; 0.5 g Polyvinylpyrrolidone; 0.5 g BSA in 50 ml); 5X SSPE (43.3 g NaCl, 6.9 g NaH2PO4.H2O and 1.85 g EDTA in 1 litre pH=7.4); 10 µg/ml *E. coli* DNA as carrier and 1%SDS]. Hybridizations were carried out with respective labeled recombinant clones (1 X 10^7 cpm of labeled probe per 10 ml of hybridization fluid) for 16 hours at 50^0C. Washings were performed at high stringency: blots were washed using 2X SSC/0.1% SDS at 65^0C twice for 15 min., followed by 1X SSC/0.1% SDS at 65^0C for 15 min. and finally 0.1X SSC/ 0.1% SDS at 65^0C twice for 15 min. Autoradiography for individual blot was done following standard procedures (Sambrook *et al.*, 1989).

3.8 RT-PCR and Southern Blotting

In order to confirm the Northern results, internal primers were designed from each uncovered fragment (Table 4) and the RT-PCR amplifications were conducted using cDNA from different tissues and the spermatozoa. The sequences and annealing temperatures of each primer set corresponding to individual fragment and amplicons' size are given in the table 4. The products were checked on 1% agarose gel in 0.5XTBE and then transferred to nylon membrane using capillary transfer method followed by pre-hybridization for 4-6 hours at 60^0C and hybridization with respective [^{32}P] α-dCTP random primed recombinant clones for 16 hours at 55^0C. Bubaline derived β-actin gene probe and bacterial genomic DNA were used as positive and negative controls, respectively. RT-PCR blots were washed at high stringency using 2X SSC/ 0.1% SDS at 65^0C twice for 15 min., followed by 1X SSC/ 0.1% SDS at 65^0C for 15 min. and finally 0.1X SSC/ 0.1% SDS at 65^0C twice for 15 min. Autoradiography was carried out as discussed in section 3.6.

Table 4: Details of the oligos used for the RT-PCR analyses for the MASA uncovered fragments #

S.No.	Oligo ID	Gene ID	Sequence (5'-3')	Annealing Temp. (In ^{O}C)
A. For the 33.15 repeat tagged transcripts				
1.	JSO199	AY762113	F 5' TTCTCCAGCTGAAAAGGACAG 3'	56.5
2.	JSO200		R 5' CCCCTTAGGAGCGGAGTT C 3'	
3.	JSO201	AY762112	F 5' TTCCCCCTTCGACATCAAGAC 3'	60.1
4.	JSO202		R 5' CGCCCAGTCCTCAGATCAT 3'	
5.	JSO203	AY762114	F 5' CAGGTGCCTCTGGCTTGT 3'	59.9
6.	JSO204		R 5' GGGCCCTTGAACTATAACCA 3'	
7.	JSO205	AY762115	F 5' CCACCACACCCTCTACCATT 3'	59.5
8.	JSO206		R 5' TTCATCTAGCTGGGCTGAGG 3'	
9.	JSO207	AY847460	F 5' AGTCCCTGGGGGCTATAATC 3'	58.0
10.	JSO208		R 5' GCCACTTGAAGCAGGTCTGT 3'	
11.	JSO254	AY920927	F 5' GCCTCCAAGTCAAGGTGAGT 3'	58.0
12.	JSO255		R 5' AGGATCCCCCAGAGAAAGAA 3'	
13.	JSO271	AY947405	F 5' GGTTTCTCATAAGTGACCGTGACC 3'	63.1
14.	JSO272		R 5' TGAGATGACCTTGTCCTTGTTCAG 3'	
B. For the GACA tagged transcripts				
1.	JSO466	DQ304116	F 5' ACTGTTTGGTGTTTCTACTCCTGAC 3'	60
2.	JSO467		R 5' GAAGACACAGATAGGGTGGTTGAT 3'	
3.	JSO468	DQ534905	F 5' CTGTTGCCTTGTAGGGTTTATACTG 3'	61
4.	JSO469		R 5' AGAAACCAGTCATCACTGTCACAAA 3'	
5.	JSO478	DQ289479	F 5' TAAGGCTTCAAAAACCAGACCA 3'	64

6.	JSO479		R 5' ACCAAGACAGAGTTCCCTCCTC 3'	
7.	JSO480	DQ494483	F 5' ACCACACATCAACTGGAAAACC 3'	64.5
8.	JSO481		R 5' CCCTGGGATTAAGACTCCAAAC 3'	
9.	JSO663	DQ534902	F 5' TCTTTCAGAGGCTAATTGGGATT 3'	62
10.	JSO664		R 5' GTTTGAGGGTTAAATGGGAGAAT 3'	
11.	JSO665	DQ534907	F 5' GGTTTCAGTTCTGACTCTGCTGT 3'	64
12.	JSO666		R 5' AATGTGAGCCCCTAAATACACAA 3'	
13.	JSO669	DQ534908	F 5' ATCGACACACTGGCACAGAC 3'	62
14.	JSO670		R 5' GACCTTTGAGACCTCGGATG 3'	
15.	JSO673	DQ534909	F 5' AGCAACACAGGGCAACATTT 3'	59
16.	JSO674		R 5' AGGACAACACGTGCAACAGT 3'	
17.	JSO675	DQ534910	F 5' TACCATTGTCCAAGTTCAAA 3'	62
18.	JSO676		R 5' AGTGGCACGTTCTGGAATTT 3'	
19.	JSO679	DQ534906	F 5' GACCTTTGAGACCTCGGATG 3'	58
20.	JSO680		R 5'ATCGACACACTGGCACAGAC 3'	
21.	JSO683	DQ534904	F 5' TGGCATGTACATCAGGTTGC 3'	58
22.	JSO684		R 5' AGGCTAGGGGTACTCCGTGT 3'	
23.	JSO687	DQ834345	F 5' GCAGGAGATGTGGGTTCAAT 3'	60
24.	JSO688		R 5' CCCAAAATCCCATATGGTTG 3'	
25.	JSO691	DQ494485	F 5' ACAAAATCACGGGTCTCGTC 3'	64
26.	JSO692		R 5' CAGGAGGTGACTACCCCAGA 3'	
27.	JSO1221	DQ834344	F 5' CCTTCTTCAGGGGATCTTCC 3'	58
28.	JSO1222		R 5' CATCGCCTTGTGTCAGAGTG 3'	
29.	JSO1223	DQ789047	F 5' ACGGCTGTCCATCTTGTTTC 3'	56.5
30.	JSO1224		R 5' AATGGCAACCCACTCCAATA 3'	
31.	JSO1225	DQ789048	F 5' ACTCACCATCCGTCCTGTTCTA 3'	58
32.	JSO1226		R 5' CCTAGGGTCCAAAACTGTGCTT 3'	
33.	JSO1227	DQ789049	F 5' GGCAGTACTGGTTCTGGATGAC 3'	59
34.	JSO1228		R 5' TTATGAGGACTCAAGCTCAGCA 3'	
35.	JSO1322	DQ834346	F 5' ACACGCAAAACCTGACTTGTGA 3'	58

36.	JSO1323		R 5' TGCCTATGGATGGTTGTTGCTA 3'	
37.	JSO1324	DQ845142	F 5' TATTGATCAGTCCTGGGGGTTC 3'	58.5
38.	JSO1325		R 5' GCTGCTACAGCCTTAGGAATCAA 3'	
39.	JSO1326	DQ845143	F 5' GAAGCATTCCTCGGATGTCA 3'	56.5
40.	JSO1327		R 5' AGACACGTTGAGGGGGAGTT 3'	
41.	JSO1328	DQ845145	F 5' ACACTCTGTGCTTAGGGCTCTG 3'	58
42.	JSO1329		R 5' TCAGAGCCAAGTCTCTCTGACC 3'	
43.	JSO1330	DQ845146	F 5' ATCCTCTTGCCCTTTTTCAGAG 3'	59.5
44.	JSO1331		R 5' ATCTGACCTGCAATCAGAGGAG 3'	
45.	JSO1360	DQ904037	F 5' AAGAAAGTGGTGGACAAGCTGA 3'	58
46.	JSO1361		R 5' CAGACACACAAAGAAGCAAGCA 3'	
47.	JSO1362	DQ904038	F 5' CAGGAGCTCAGCTGGTAAAGAA 3'	56
48.	JSO1363		R 5' CACGAGTGTGGTCTTTGGAAAT 3'	
49.	JSO1364	DQ904039	F 5' TGCATTTGCTGTGATTTAGGTG 3'	58
50.	JSO1365		R 5' TCTTCGGGAGTCAAACAGGATA 3'	
51.	JSO1366	DQ913640	F 5' CAGACAACTGCAACATCACCAT 3'	60
52.	JSO1367		R 5' TTCTGTTCCTGAGTCCCACTTC 3'	
53.	JSO1368	DQ913641	F 5' GAAACAGGGAGAAAGACAAGCA 3'	58
54.	JSO1369		R 5' CAGCCACACACAGAGGTACTGA 3'	
55.	JSO1370	DQ913642	F 5' GCAACCCACTCCAGTGTTCTTA 3'	57
56.	JSO1371		R 5' GAGTGGGAGGTGTTGTGTCAGT 3'	
57.	JSO1372	DQ913644	F 5' CCGTGAGTTACTGATGGACAGG 3'	59
58.	JSO1373		R 5' GACAGACATCAAAAGGCTGACC 3'	
59.	JSO1374	DQ913645	F 5' CAGACAGACACTTGGGGCTA 3'	57
60.	JSO1375		R 5' TGGTGGCTTGTGTCTCACTC 3'	
61.	JSO1376	DQ913646	F 5' AACCCCCACCTCAAGGAGTA 3'	60
62.	JSO1377		R 5' TATGGCCCTTTCTGTTCCTG 3'	
63.	JSO1070	BETA-ACTIN	F 5' CAGATCATGTTCGAGACTTCAA 3'	59

64.	JSO1071		R 5' GATGATCTTCATTGTGCT 3'	
C. For the GATA tagged transcripts				
1.	JSO1378	EF050082	F 5' AATCACCACTTTGCAACCACACT 3'	58
2.	JSO1379		R 5' CTTCATACCCAGATGCAGACGAT 3'	
3.	JSO1385	EF050084	F 5' TAGCAGCAGAAATGGACTCAACC 3'	56
4.	JSO1386		R 5' ATACCCAGGGATTGGCACATAGT 3'	
5.	JSO1387	EF051516	F 5' TTTTGTGTCCTGCAGTTCTCTGA 3'	56
6.	JSO1388		R 5' CCTTTGTTCTGGGTACGGTGTAA 3'	
7.	JSO1389	EF051517	F 5' ATCTTGCTTTCTGGAAGACCATC 3'	55
8.	JSO1390		R 5' CAGTGGAAAATCTGTGTGCAATG 3'	
9.	JSO1391	EF051518	F 5' TTAAGCAAGACCCATTCTGTTGC 3'	57
10.	JSO1392		R 5' CAGGAAACATGTCTTTGGTTTTACA 3'	
11.	JSO1393	EF051519	F 5' TTAAGCAAGACCCATTCTGTTGC 3'	55
12.	JSO1394		R 5' CAGGAAACATGTCTTTGGTTTTACA 3'	
13.	JSO1395	EF592582	F 5' GCTACACCACCTTCATGGTCAAC 3'	57
14.	JSO1396		R 5' CATGTGTGTGTGTGTGTGTGTGT 3'	
15.	JSO1397	EF592583	F 5' GATAGATATGGGCTTCCCTGGTG 3'	58
16.	JSO1398		R 5' GCAAGAATACTGGAGTGCGTTTC 3'	

Note that primer IDs alongwith their respective gene accession numbers are also given in the table.

3.9 Evolutionary studies of the uncovered genes/gene fragments

The evolutionary status of the uncovered genes/gene fragments was studied using them as the probes in individual southern blots with the DNA from various mammal and non-mammal species. Approximately, 500 ng of heat denatured genomic DNA from above mentioned species was briefly run on 0.8% agarose gel and transferred onto the nylon membrane using capillary transfer method. Hybridizations were conducted with individual labeled probes following the nick translation method. Phylogenetic tree were constructed for each uncovered gene/gene fragments using different online available softwares for e.g. clustalW (http://www.ebi.ac.uk/clustalw.html) and PHYLIP (http://evolution.genetics.washington.edu/phylip.html).

3.10 Copy number calculation and Relative expressional studies using Real Time PCR

Copy number of the MASA uncovered genes and *c-kit* was calculated based on absolute quantitation assay using SYBR Green dye and Sequence Detection System-7000 (ABI, USA). For every gene, the standard curve was obtained using 10 fold dilution series (From 300 copies to 30 million copies) of the recombinant plasmids containing gene of interest. Standard curve with a slope of -3.2 and a single dissociation peak substantiating maximum efficiency of the PCR reaction and high specificity of the primers with target DNA respectively, was obtained. Similar standard curve analysis was performed using cDNA and genomic DNA as templates before performing the actual experiments. The copy number of the respective genes/fragments was calculated by extrapolating the standard curves using calculated copies of the haploid buffalo genome as the target template in the Real Time PCR absolute quantitation assays. For relative expression analysis, SYBR green assays were conducted for individual fragments using equal amount of cDNA from all the tissues and spermatozoa, with β-actin as an internal control. The expression level or amount of transcripts of a particular fragment was estimated by comparative Ct method with one of the tissues as calibrator sample. Primers for calculating copy number and relative expression for each

transcript were designed by "Primer Express Software" (ABI, USA) and are given in the table 5. The cyclic conditions comprised 10 minutes of polymerase activation at 95^0C followed by 40 cycles, each at 95^0C for 15 seconds and 60^0C for 1 minute. Presence of genomic DNA within the cDNA was ruled out by using mRNA from different tissues as template in several independent Real Time PCR reactions. Further, a no template control was run to ensure that the samples with lowest detected cDNA amount are well above the background signal. Each experiment was repeated three times at different concentration to ensure consistency of the results. The expression level of the genes was calculated using the formula: expression status= $(1+E)^{-\Delta Ct}$, where E is the efficiency of the PCR and ΔCt is the difference between cycle threshold of the test sample and endogenous control.

Copy number of *Smoc*-1 gene was calculated using two primer pairs (Table 3) specific to 11^{th} exon of *Smoc*-1. The relative transcription of *Smoc*-1 across the tissues was assessed using the same primer set and equal quantity of total cDNA from different tissues and semen samples of buffalo. The relative expression of both the detected transcript variants was studied using three primer sets, two for both the variants (JSR1015-1018) and one specific for variant-01 (JSR1033-1034). Age specific expression of both the transcripts was also carried out using cDNA from blood lymphocytes with the primer sets (JSR1015-1018) which picked up both the variants and another one JSR1033-1034 specific for variant-01. The total cDNA amounts from different tissues were optimized using a set of buffalo β-actin primers (forward 5'TCACGGAGCGTGGCTACAG3' and reverse 5'TTGATGTCACGGA CGATTTCC3'). Presence of genomic DNA in the cDNA templates was ruled out by using mRNA as template in several independent Real Time PCR and end point PCR reactions. Each reaction was repeated three times in triplicates.

3.11 Metaphase chromosome preparation and Fluorescent in situ hybridization (FISH)

Approximately, 400 µl of whole blood from normal buffaloes was cultured for chromosome preparations in the RPMI-1640 media containing

Table 5: Details of the oligos used for the Copy number calculation and Relative expressional studies for the MASA uncovered genes/ gene fragments #

S.N.	Oligo ID	Gene ID	Sequence (5'-3')	Annealing Temp. (In ^{O}C)
A. For the 33.15-taagged transcripts				
1.	JSR213	AY847460	F 5' TTATGTGGGTGGAATAACCTAGTTTG 3'	60
2.	JSR214		R 5' GCCATACTCTGGCCCCTAACA 3'	
3.	JSR269	AY947405	F 5' TTCAGGTTCGGTCACCGATAA 3'	60
4.	JSR270		R 5' TCGGCTTAGACCCGTCATCT 3'	
5.	JSR215	AY762114	F 5' CCAGTCTCCCCCGCCTAT 3'	60
6.	JSR216		R 5' AGGCAGGGATACCAAAATGCT 3'	
7.	JSR294	AY762112	F 5' TTCGACATCAAGACCCTCATCA 3'	60
8.	JSR295		R 5' ACGACGCCCAGGAAGGA 3'	
9.	JSR298	AY920927	F 5' GACTGAAGAAAGACAGCAAGATTGTAA 3'	60
10.	JSR299		R 5' ACCCAGGCTTAGAGAGAAAGCA 3'	
11.	JSR296	AY762115	F 5' TGCATGCCAGATAGAGACAGAAC 3'	60
12.	JSR297		R 5' GGGCTATCCCCCTGACCTT 3'	
13.	JSRH01	AY762113	F 5' CCGGGCTTCTCAGAGGTA 3'	60
14.	JSRH02		R 5' AGCGGAGTTCGGCTGTCA 3'	
B. For the GACA identified transcripts				
1.	JSR472	DQ304116	F 5' TGAGTGGGAGGAGGAGAAATACTT 3'	60
2.	JSR473		R 5' CTACCCCAGCCCGGGTTA 3'	
3.	JSR474	DQ534905	F 5' GTGAGTTCGTCTCTGGAAACCAT 3'	60
4.	JSR475		R 5' CAGACAATCTTCCCTGCTTTCTG 3'	

5.	JSR482	DQ494483	F 5' GTGGCCCTGTTTGAAAAACC 3'	60
6.	JSR483		R 5' TGCGGGAGAGTCTTCCAGAT 3'	
7.	JSR484	DQ289479	F 5' CACCCCTGCAGCTGATGAA 3'	60
8.	JSR485		R 5' GGACTGTCCACTTGCCTTCCT 3'	
9.	JSR605	DQ534907	F 5' TGGGTCACTCTTCTGCTTCTAACA 3'	60
10.	JSR606		R 5' CTGCAGAGTCTTTGAGAATTTTGG 3'	
11.	JSR607	DQ534902	F 5' AGCGCGGTACTGCGTGTAA 3'	60
12.	JSR608		R 5' TTGCCCTGTGTCAACCGATA 3'	
13.	JSR611	DQ534910	F 5' AAGGGTAGGATGCCCATTGAG 3'	60
14.	JSR612		R 5' GCACGTTCTGGAATTTCCTAAAG 3'	
15.	JSR615	DQ534906	F 5' TCCCTTAGGGAAGCTGCTTCT 3'	60
16.	JSR616		R 5' CGCTCCGAGATCGACACACACT 3'	
17.	JSR697	DQ834344	F 5' TGCCAGGCTCCTCTGTCACT 3'	60
18.	JSR698		R 5' AAGATCCCCTGAAGAAGGAAATG 3'	
19.	JSR699	DQ494485	F 5' TGTAACCGGCCAGTGACTCA 3'	60
20.	JSR700		R 5' CTCCCTGTGATGCCAGCTTT 3'	
21.	JSR701	DQ534904	F 5' CTGAGCATGCAGCCTGTAGGT 3'	60
22.	JSR702		R 5' GTGCAGGCAGGTGTCTAAAGG 3'	
23.	JSR703	DQ534909	F 5' TTAAAAAACACACCTGAGTTGAAAAGTG 3'	60
24.	JSR704		R 5' ATGTGAAGTGCAAGCCTATTTTAGG 3'	
25.	JSR705	DQ534908	F 5' CATCCGCTCCCAGATCGA 3'	60
26.	JSR706		R 5' CCTCCCTTAGGGAAGCTGCTT 3'	
27.	JSR707	DQ834345	F 5' TGGCAACCCACTCCCGTAT 3'	60
28.	JSR602		R 5' CATGTCCGACTTTTTGTGACCTT 3'	
29.	JSR619	DQ913644	F 5' TGAGCGACTGAACTGAGGTGAA 3'	60
30.	JSR620		R 5' TGACCGTGAGCCCCAGAA 3'	
31.	JSR625	DQ913645	F 5' TGTGCCTGAGCAGACATGTGT 3'	60
32.	JSR626		R 5' TGTTTGCTTTTCTTTGCCTACAAT 3'	
33.	JSR631	DQ904039	F 5' TGATTTAGGTGGGAGCCTGAGA 3'	60
34.	JSR632		R 5' GTGCCGTCACCCCTAGGA 3'	
35.	JSR635	DQ904036	F 5' GTGAGTTCGTCTCTGGAAACCAT 3'	60

36.	JSR636		R 5' CAGACAATCTTCCCTGCTTTCTG 3'	
37.	JSR639	DQ789048	F 5' ACGAGGAGCCTTGTTTGTCTAC 3'	60
38.	JSR640		R 5' TGCCCACGTGACTCTTTCTG 3'	
39.	JSR643	DQ789047	F 5' GCACATTCAAGGGCTCATTCATCA 3'	60
40.	JSR644		R 5' GGCTGCATAGTGTCCCTGTCT 3'	
41.	JSR818	DQ789046	F 5' CGAGGAGCCTTGTTTGTCTACA 3'	60
42.	JSR819		R 5' TGCCCACGTGACTCTTTCTG 3'	
43.	JSR649	DQ845146	F 5' TCCTCTTGCCCTTTTTCAGAGT 3'	60
44.	JSR650		R 5' ACACCAGTCATCAGGTATGACATTG 3'	
45.	JSR653	DQ845144	F 5' TCCCTTAGGGAAGCTGCTTCT 3'	60
46.	JSR654		R 5' GAGATCGACACACTGGCACAGA 3'	
47.	JSR661	DQ916743	F 5' CACCCCCCGTGACTTCCT 3'	60
48.	JSR662		R 5' GCCTGTTGGGCTCCCTTT 3'	
49.	JSR712	DQ904038	F 5' GTATTCTCGCCTGGAAATGC 3'	60
50.	JSR711		R 5' CCAACTCTTATAGGACCCATGGA 3'	
51.	JSR714	DQ845143	F 5' GAAGGAACCAGAGCTACTCACAGTCTA 3'	60
52.	JSR715		R 5' GGGCACCGAGGAGTCATG 3'	
53.	JSR716	DQ845142	F 5' GCATAAGTTTGAGCAAGCTTTGG 3'	60
54.	JSR717		R 5' TACGACGTCAGGTACCCCA 3'	
55.	JSR718	DQ789049	F 5' TCGGCATAAGCTGGGAGTAAT 3'	60
56.	JSR719		R 5' CACCTGATCGCCATTTCCA 3'	
57.	JSR720	DQ834346	F 5' CATGGTACACGTGAAAGAATGACTTC 3'	60
58.	JSR721		R 5' TTCCAATCTTTGCTTTGCATTAAATC 3'	
59.	JSR724	DQ834347	F 5' CATCCGCTCCGAGATCGA 3'	60
60.	JSR725		R 5' CCTCCCTTAGGGAAGCTGCTT 3'	
61.	JSR726	DQ913641	F 5' CTGGCCTGGTGTTTATGTACGA 3'	60
62.	JSR727		R 5' CAGAAAAGGACTAGATACACGGAAG 3'	
63.	JSR728	DQ913640	F 5' GACAACTGCAACATCACCATTC 3'	60
64.	JSR729		R 5' AAAGTGCAGCTTTCCAGTCTTGTA 3'	
66.	JSR730	DQ845141	F 5' CCTGCGCACAGCAGACATA 3'	60

	JSR731		R 5' CAGAGTAACACAGCAGAGTCAGAACTG 3'	
67.	JSR732	DQ904037	F 5' TGCCATGAAGGTTCTAAGAAAGTG 3'	60
68.	JSR733		R 5' TGGTTCTGGGTCTCTCTCTAACTGTACA 3'	
69.	JSR734	DQ789045	F 5' TGCAACAAGTCACCTGAAAAGATT 3'	60
70.	JSR735		R 5' GGCCCCGCCGTCTCT 3'	
71.	JSR814	DQ913646	F 5' CAAGACTCTGAGGTGCAAATGG 3'	60
72.	JSR815		R 5' CTAAGCCCTCAGGCCCCTAT 3'	
73.	JSR627	DQ913642	F 5' AGTCAGCTGGAGAAGGAAAAGGAAATGG 3'	60
74.	JSR628		R 5' TGTGACCCCGTGGACTGTAG 3'	
75.	JSR1399	DQ845145	F 5' AGGGAAGGCTTCTCAGAGGAA 3'	60
76.	JSR1400		R 5' TGAGAGAGTGGGAAGGGTCTTTC 3'	
C. For the GATA uncovered transcripts				
1.	JSR876	EF050082	F 5' ATCACCACTTTGCAACCACACT 3'	60
2.	JSR877		R 5' TTGCCCGCTAGTTTCACATTC 3'	
3.	JSR822	EF050084	F 5' ACTATGTGCCAATCCCTGGGTAT 3'	60
4.	JSR823		R 5' TTCCTGAGCTCGTCACTCCAT 3'	
5.	JSR878	EF051516	F 5' GCACTTTCCTTTTGTTATAGTAGCTTGTAG TATTC 3'	60
6.	JSR879		R 5' TGTTCTGGGTATGCTTAAACCA 3'	
7.	JSR826	EF051517	F 5' AGCAGCAGAAATAACTATAGGGTATCCT 3'	60
8.	JSR827		R 5' TGTATGCTGTATCACAGTGGAAAATC 3'	
9.	JSR883	EF051518	F 5' TGCATCTGACCTTTAAGAACAGAATT 3'	60
10.	JSR884		R 5' AAGACTCCACCTACAATGTACTTTCACT 3'	
11.	JSR1084	EF592582	F 5' TGGCAGAGGCAGTCTGTATCC 3'	60
12.	JSR1085		R 5' GTGTAGCTGCTTGGCTTGCA 3'	
13.	JSR1087	EF592585	F 5' CCTACAATGTACTTTCACTTTCAATCATC 3'	60

14.	JSR1088		R 5' GCATCTGACCTTTAAGAACAGAATTTATG 3'	
15.	JSR1089	EF592583	F 5' AATATGGGAGATCTGGGTTCGA 3'	60
16.	JSR1090		R 5' TGGAGTGCGTTTCCATGCT 3'	
19.	JSR522	BETA-ACTIN	F 5' TCACGGAGCGTGGCTACAG 3'	60
20.	JSR523		R 5' TGGATGTCACGGACGATTTCC 3'	

Note that primer IDs alongwith their respective gene accession numbers are also given in the table.

20% fetal bovine serum for 72 hours at 37^0C with 5% CO_2. Then, the colcemid was added to the culture according to supplier's specifications in order to arrest the chromosome at metaphase stage. The cells were treated using hypotonic treatment (0.56% KCl) followed by fixation (3:1 ratio of methanol: acetic acid) thrice. Cattle derived BAC clone Ctg9.CH240-54I18 representing full length *Smoc*-1 gene and human derived BAC clone RP11-571F15 for Ubap1 gene were used as probes. All the BAC clones were purchased from Children's Hospital Oakland Research Institute (CHORI). ANKD26 was localized using recombinant plasmid containing insert of 523 bp (representing the part of Ankyrin repeat domain-26). Probes were labeled with Fluorescein-12-dUTP using Nick Translation Kit from Vysis, (IL, USA) and detected with biotinylated anti-fluorescein antibody and FITC-avidin DCS (Vector Labs). Two rounds of signal amplifications were performed to obtain the desired signal intensity. Washing, counterstaining, and mounting of the slides were conducted following supplier's instructions. The slides were screened under the Olympus Fluorescence Microscope (BX51) fitted with vertical fluorescence illuminator U-LH100HG UV, excitation and barrier filters. Images were captured with a CCD camera. Chromosome identification and band numbering were done through G-banding following the International System for Chromosome Nomenclature of Domestic Bovids (ISCNDB, 2000).

3.12 Protein expression and production of anti-Smoc-1 antiserum

Using PSmoc-1 as template, the Smoc-1 was re-amplified to accommodate a *Bam*HI site at the 5' end (5'-CGGGATCCCACCTC TCCACCTGCCCCAGG-3') and *Xho*I site at the 3' end (5'-CCCTCGAGTTAGACG AGGCGTCCTACTTC-3'). The resulting amplicon was cloned in pGEX-4T1 vector (Novagen, USA) at *BamH*I/ *Xho*I sites. The recombinant GST-tag-Smoc1 was transformed in to BL21 (DE3) *E.coli* and expression of the recombinant protein was induced with 1mM IPTG at 25 to 37^0C for 4 hour. The recombinant Smoc-1 protein was purified using GST-tag purification resin (Clontech, USA). A rabbit was immunized with purified recombinant pGEX-4T1-PSmoc1 using alum as an adjuvant to

obtain the Anti-PSmoc1-pAb. To ensure the specificity, primary antiserum (Anti-SySmoc1-pAb) was obtained for a commercially synthesized 26 amino acid (69S to 95G) long peptide, conjugated to Keyhole limpet hemocyanin (KLH), specific to Smoc-1 domain.

3.13 Isolation of total protein from different tissues and Western Blotting

Denaturing 10% polyacrylamide gels were used under reducing conditions for analyzing culture medium and *E. coli* expressed proteins. Following electrophoresis, the proteins were transferred onto nitrocellulose membranes. After blocking with 5% non-fat milk, 1% BSA in PBS for 45 min at room temperature, the membrane was probed with primary antibodies (anti-PSmoc1-pAb raised against pGEX-4T1-PSmoc1 anti-SySmoc1-pAb against synthesized peptide of Smoc-1). Secondary detection was carried out with goat anti-rabbit IgG conjugated with HRP (Bio-rad, USA) following standard protocol. Total protein was isolated from different tissues using Tri-X reagent (MRC) followed by extraction of lower and middle layers. Protein was precipitated from these layers using acetone. Protein quantity was normalized on 12% SDS-PAGE followed by its transfer, hybridization and detection.

3.14 Immunohistochemical analysis of Smoc-1 on Buffalo Tissue Sections

The distribution of Smoc-1 protein in different tissues was studied on paraffin sections by indirect Immunohistochemistry using Anti-SySmoc1-pAb (generated against synthesized polypeptide for Smoc-1 as stated earlier). Freshly prepared buffalo tissues were fixed for 1 hr in 4% Paraformaldehyde/PBS and after dehydration were embedded in paraffin. Sections were sliced and processed according to standard procedures (Sambrook *et al.*, 1989). After blocking with 1% BSA/TBS, the sections were incubated with the anti-PSmoc1-pAb followed by HRP-labeled goat anti-rabbit IgG (Bio-rad, USA) and detected using Di-amino Benzene (DAB) as substrate. The sections were observed under BX-51 microscope (Olympus, JAPAN).

RESULTS

4. RESULTS

4.1 MASA mediated mining of the mRNA transcripts

4.1.1 Distribution of GACA/GATA repeats and consensus of 33.15 repeat loci across the species

Before conducting the MASA experiments, a comprehensive database search was done for the distributional analysis of GACA/GATA and 33.15 repeat sequences in the non-coding and coding genomes across the species (Figure 4). In case of buffalo, the slot blot hybridization of the total RNA from different tissues of buffalo with oligos based on the GACA, GATA and 33.15 repeats showed discernible but differential signals in all the tissues indicating tagging of these repeats with several transcripts (Figure 5). The presence of all the repeat sequences across the species (*in silico*) and various buffalo tissues (demonstrated experimentally) is given below.

4.1.1.1 *Repeat motifs within the non-coding genomes across the species*

The detailed *in-silico* analysis of the available complete or incomplete genomes of Archea/ Eubacteria (Prokaryotes) and 17 eukaryotes including human revealed absence of GACA/GATA tetramers in the prokaryotes (Table 6). Lower eukaryotes like *Saccharomyces cervisae*, *Dictyostelium discoideum*, *C. elegans*, and *Arabidopsis thaliana* also harbored either no or very few repeats. However, a gradual accumulation of these repeats was observed in the higher eukaryotes (Table 6). Thereafter, detailed analysis of 6 of the above mentioned eukaryotic species evidenced the differential occurrence of the GACA/GATA repeats among different chromosomes and species (Figure 4). Of these, the human, dog and *Arabidopsis* genomes were found to be GATA rich and chicken with equal frequency of GAC/GATA repeats, whereas cattle remained indecipherable due to its unfinished genome. The *C. elegans* genome was found to harbor only 13 regions containing $(GACA)_4$ and 12, $(GATA)_4$ repeats. When considered individually, the highest frequency of GACA was found

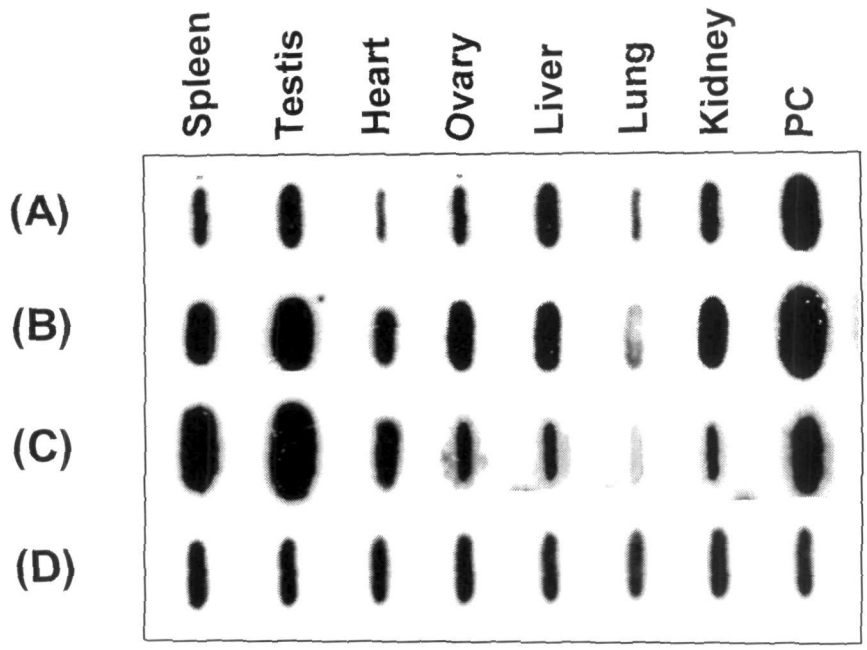


Figure 5. Slot blot hybridization of 5 µg total RNA with different oligo probes showing varying signal in different somatic tissues and gonads of buffalo. The probe labeled for the consensus of 33.15 repeat loci **(A)**; (GACA)4 in **(B)** and for (GATA)4 has been shown in **(C).** The normalized signal of RNA with β-actin in all the tissues has been given in **(D)**. Uniform signal in panel "**D**" indicates similar quantity of the RNA used. PC denotes 5 pmole of the respective oligos used as positive control.

Table 8: Occurrences of the *Bkm* derived GATA repeats in the coding genes across the species#

S.N.	Encoding genes/Transcripts	Species	Accession no.	Length	Repeat position
1.	Sorting nexin 1 (SNX1)	*Homo sapiens*	NM_152826	8077	4827-4869
2.	Ras association (RalGDS/AF-6) and pleckstrin homology domains 1 (RAPH1)	*Homo sapiens*	NM_213589	9615	9356-9374
3.	Ankyrin repeat domain 5 (ANKRD5)	*Homo sapiens*	NM_198798	3775	3463-3536
4.	Peripheral myelin protein 2 (PMP2)	*Homo sapiens*	NM_002677	3579	2330-2372
5.	NK2 transcription factor related, locus 3	*Homo sapiens*	NM_145285	2083	1883-1903
6.	Potassium channel, subfamily K, member 2 (KCNK2)	*Homo sapiens*	NM_001017425	3212	2168-2217
7.	Coiled-coil and C2 domain containing 1B (CC2D1B)	*Homo sapiens*	NM_032449	5639	2913-2933
8.	Origin recognition complex, subunit 6 like (yeast) (ORC6L)	*Homo sapiens*	NM_014321	1647	1395-1449
9.	Glycerophosphodiester phosphordie-sterase domain containing 4 (GDPD4)	*Homo sapiens*	NM_182833	2545	1806-1824
10.	TBP-associated factor 9L	*Homo sapiens*	XM_925841	6660	3192-3232
11.	Kynurenine 3-monooxygenase (kynurenine 3-hydroxylase)	*Homo sapiens*	NM_003679	4992	2086-2152
12.	Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 1 (SLC11A1)	*Homo sapiens*	NM_000578	3865	2402-2422
13.	Regulator of G-protein signalling 5 (RGS5)	*Homo sapiens*	NM_003617	5848	3144-3158
14.	WD repeat domain 72 (WDR72)	*Homo sapiens*	NM_182758	5900	4373-4388
15.	Zinc finger and BTB domain containing 4 (ZBTB4)	*Homo sapiens*	NM_020899	5888	5245-5259
16.	Sialic acid binding Ig-like lectin 1, sialoadhesin (Siglec1)	*Mus musculus*	NM_011426	6875	6818-6872

17.	Simple repeat sequence-containing transcript (Srst)	*Mus musculus*	NM_009276	1263	1080-1132
18.	Extracellular matrix protein 2	*Mus musculus*	NM_001012324	3627	3243-3275
19.	SH3-domain GRB2-like 2 (Sh3gl2)	*Mus musculus*	NM_019535	2817	1724-1784 1833-1861 1972-2008
20.	Bone morphogenetic protein 8b (Bmp8b)	*Mus musculus*	NM_007559	2942	1940-2008
21.	Choline kinase alpha (Chka)	*Mus musculus*	NM_001025566	3644	1414-1446
22.	Melanocortin 3 receptor (Mc3r)	*Mus musculus*	NM_008561	2624	205-237
23.	Neurotrophic tyrosine kinase, receptor, type 2 (Ntrk2)	*Mus musculus*	NM_008745	7022	2533-2585 2662-2688
24.	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3 (Pfkfb3)	*Mus musculus*	NM_133232	4991	3515-3565
25.	N-acylsphingosine amidohydrolase (acid ceramidase)-like (Asahl)	*Mus musculus*	NM_025972	1827	1525-1568
26.	Prokineticin receptor 2 (Prokr2)	*Mus musculus*	NM_144944	3655	1767-1786
27	NACHT, LRR and PYD containing protein 12, transcript variant 1 (Nalp12)	*Mus musculus*	XM_973733	3169	2796-2828
28.	Protein phosphatase 1, regulatory (inhibitor) subunit 3E (Ppp1r3e)	*Mus musculus*	XM_989516	3996	350-400
29.	Similar to 60S ribosomal protein L29	*Mus musculus*	XM_975825	1904	33-79
30	RAB11 family interacting protein 1	*Mus musculus*	XM_922410	2326	1980-2006
31.	Myosin VC, transcript variant 9 (Myo5c)	*Mus musculus*	XM_925601	2253	1498-1554
32.	Sorbin and SH3 domain containing 2, transcript variant 1 (Sorbs2)	*Mus musculus*	XM_989416	892	634-698
33.	Nidogen 1 (Nid1)	*Rattus norvegicus*	XM_213954	5220	5180-5220
	Solute carrier family 10, member 2 (Slc10a2)	*Rattus norvegicus*	NM_017222	4269	1725-1779
34.	Myotubularin related protein 2	*Rattus norvegicus*	XM_001068538	3668	3449-3515
35.	Neurotrophic tyrosine kinase, receptor, type 2 (Ntrk2)	*Rattus norvegicus*	NM_012731	4797	3404-3422

36.	Solute carrier organic anion transporter family, member 1b2 (Slco1b2)	*Rattus norvegicus*	NM_031650	3212	2778-2918 2951-3005
37.	Cytochrome P450, subfamily 11B, polypeptide 1 (Cyp11b1)	*Rattus norvegicus*	NM_012537	2696	1944-1984
38.	C-reactive protein, pentraxin-related (Crp)	*Rattus norvegicus*	NM_017096	1678	1216-1310
39.	Neurotrophic tyrosine kinase, receptor, type 3 isoform b	*Bos taurus*	XM_585006	1689	1533-1547
40.	LIM and senescent cell antigen-like domains 1	*Gallus gallus*	XM_423884	4725	259-274
41.	Ran-binding protein 2	*Gallus gallus*	XM_423497	1165	2-17
42.	Calpain inhibitor (Calpastatin)	*Gallus gallus*	XM_424713	3551	348-365
43.	Synuclein, beta (SNCB)	*Gallus gallus*	NM_204671	1253	998-1018
44.	Iduronidase, alpha-L- (IDUA)	*Gallus gallus*	NM_001031433	4906	522-537
45.	Poly(rC)-binding protein 3 (Alpha-CP3), transcript variant 2	*Danio rerio*	XM_703071	1177	1053-1161
46.	Monoamine oxidase (mao)	*Danio rerio*	NM_212827	4456	2604-2780
47.	Forkhead box O5 (foxo5)	*Danio rerio*	NM_131085	4547	72-88
48.	Protocadherin 1 gamma b 2 (pcdh1gb2)	*Danio rerio*	NM_001012658	4332	4129-4147
49.	Tumor protein D52-like 2 (tpd52l2	*Danio rerio*	NM_199582	2259	1064-1078
50.	ADP-ribosylation factor 3a (arf3a)	*Danio rerio*	NM_001003441	1475	798-814
51.	CASP2 and RIPK1 domain containing adaptor with death domain (cradd)	*Xenopus tropicalis*	NM_001006910	1711	1000-1032
52.	Phosphodiesterase 6 CG8279-RA (Pde6),	*Drosophila melanogaster*	NM_142112	5136	4076-4099
53.	CXIP1 (CAX INTERACTING PROTEIN 1)	*Arabidopsis thaliana*	NM_115347	806	589-604
54.	NADK1; NAD+ kinase (NADK1)	*Arabidopsis thaliana*	NM_113001	2323	1754-1768
55.	ATP binding / kinase/ protein kinase/ protein serine/threonine kinase/protein-tyrosine kinase (AT4G31170)	*Arabidopsis thaliana*	NM_001036681	1735	1543-1557

56.	Knotted1-like homeodomain protein liguleless3 (lg3)	*Zea Mays*	AF457124	912	713-729
57.	Glutamine synthetase (gs1-2)	*Zea Mays*	AF359511	1505	1345-1361
58.	Ramosa 2 (ra2)	*Zea Mays*	DQ327701	6032	3742-3762
59.	Cinnamoyl-CoA reductase (ccr2)	*Zea Mays*	AY227034	1239	1181-1198

Some species such as Archeas, *Sus scrofa*, *Ovis aries*, *C. familiaris*, *C. elegans* and *D. discoideum* are devoid of this repeat.

Table 9: Occurrences of consensus repeat of 33.15 loci in the different mRNA transcripts across the species

S.No	mRNA transcripts	Length	Species	Nucleotide position	Homology
1.	Pottasium channel, subfamily K	794	*Bos taurus*	460-473	14/14
2.	ß-galactosamide alpha-2, 6-sialyltransferase (SIAT7B)	1427	*Bos taurus*	1271-1284	14/14
3.	Gap junction protein, beta 3 (GJB3), transcript variant 2	1777	*Homo sapiens*	1029-1043	15/15
4.	Gap junction protein, beta 3 (GJB3), transcript variant 1	2220	*Homo sapiens*	1486-1500	15/15
5.	Neurexin 2 (Nrxn2), transcript variant alpha 1	6616	*Homo sapiens*	3699-3712	14/14
6.	Pottasium channel, subfamily K, member 18 (KCNK18)	1155	*Homo sapiens*	822-835	14/14
7.	Centaurin, alpha 2 (CENTA2)	2578	*Homo sapiens*	34-47	14/14
8.	TP53 dependent G2 arrest mediator candidate	1496	*Homo sapiens*	912-925	14/14
9.	Syntaphilin (SNPH)	5026	*Homo sapiens*	290-303	14/14
10.	Spectrin repeat containing, nuclear envelope 1	27652	*Homo sapiens*	18223-18236	14/14
11.	Neurexin 2 (Nrxn2), transcript variant alpha 2	6390	*Homo sapiens*	3579-3592	14/14
12.	Adenergic, alpha-1A-receptor (ADRA1A), transcript variant2	2306	*Homo sapiens*	172-185	14/14
13.	Adenergic, alpha-1A-receptor (ADRA1A), transcript variant3	2089	*Homo sapiens*	172-185	14/14
14.	Disintegrin and metalloproteinase domain II (ADANII)	2908	*Homo sapiens*	174-187	14/14
15.	Zinc finger protein 36, C3Htype-like 1 (ZFT36L1)	3022	*Homo sapiens*	1712-1725	14/14
16.	ATP-binding cassette, sub-family B (MDR/TAP), member 10	3460	*Rattus norvegicus*	1768-1781	14/14
17.	Neurexin 2 (Nrxn2)	5436	*Rattus*	3485-3498	14/14

			norvegicus		
18.	ATPase, Ca++ transporting, plasma membrane 3 (Atp2b3)	4470	*Rattus norvegicus*	1241-1254	14/14
19.	ATP binding cassette sub family B (NDR/TAP)	2574	*Rattus norvegicus*	2123-2136	14/14
20.	Syntaphilin (Snph) protein	4689	*Mus musculus*	385-398	14/14
21.	E2F transcription factor 7	5434	*Mus musculus*	3210-3223	14/14
22.	Anterior homebox containing gene 2 (Vax2)	1240	*Mus musculus*	1032-1045	14/14
23.	Ligatin (Lgtm)	2094	*Mus musculus*	683-696	14/14
24.	Myocin VI (Myo6)	4602	*Mus musculus*	3990-4003	14/14
25.	Neurexin II (Nrxn2)	5120	*Mus musculus*	3110-3123	14/14
26	Neurexin II (Nrxn2)	5656	*Mus musculus*	3272-3285	14/14
27.	ATPase, Ca++ transporting, plasma membrane 3 (Atp2b3)	4483	*Mus musculus*	1248-1261	14/14
28.	GATA5 protein (LOC485961)	3093	*Canis familiaris*	1070-1083	14/14
29	Mkiaa0921 protein (LOC483760)	7809	*Canis familiaris*	5387-5400	14/14
30.	Polydom protein (LOC474699)	10667	*Canis familiaris*	876-889	14/14
31.	RNA binding protein	9594	*Canis familiaris*	5861-5874	14/14
32.	ACOS3 protein	3165	*Canis familiaris*	2933-2946	14/14
33.	Ankyrin repeat domain 27 (VPS domain)	4542	*Canis familiaris*	2837-2850	14/14
34.	Sodium/glucose cotransporter (SGLT1)	2013	*Sus scrofa*	517-531	15/15

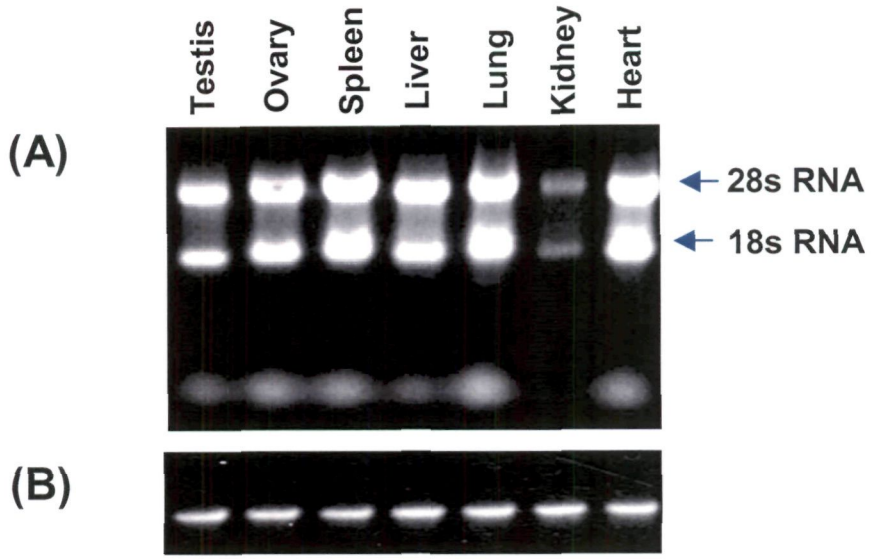


Figure 6. A representative gel showing total RNA isolated from different tissues of water buffalo **(A)**. The quantity of each RNA sample was spectrophotometerically measured and 5 μg of total RNA was reverse transcribed into cDNA. The cDNA was normalized by PCR amplifications using primers for the β-actin **(B)**.

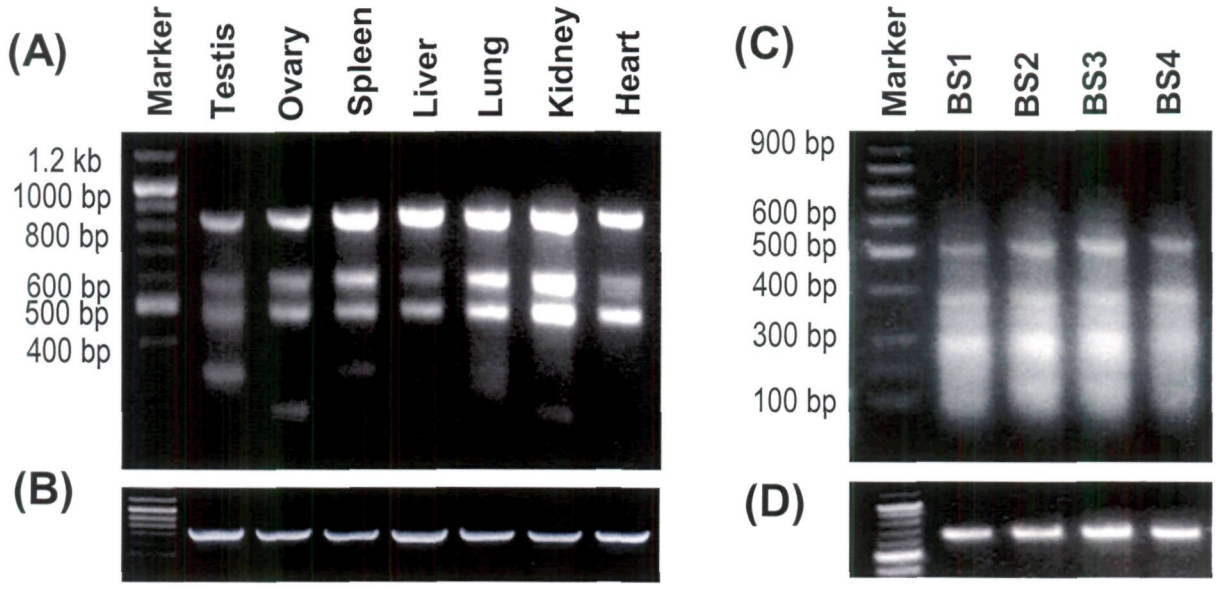


Figure 7. Minisatellite Associated Sequence Amplification (MASA) with cDNA from different tissues of buffalo using 16 base long oligo primer 33.15 showing 3-5 bands in each lane **(A)** and PCR amplification of cDNA with β-actin primers derived from buffalo showing almost uniform expression used for normalization **(B)**. A 1263 bp fragment from liver was also detected in another MASA reaction (not shown here). MASA reactions were performed using cDNA from the buffalo spermatozoa **(C)**, and the respective cDNAs were normalized by the β-actin **(D)**.

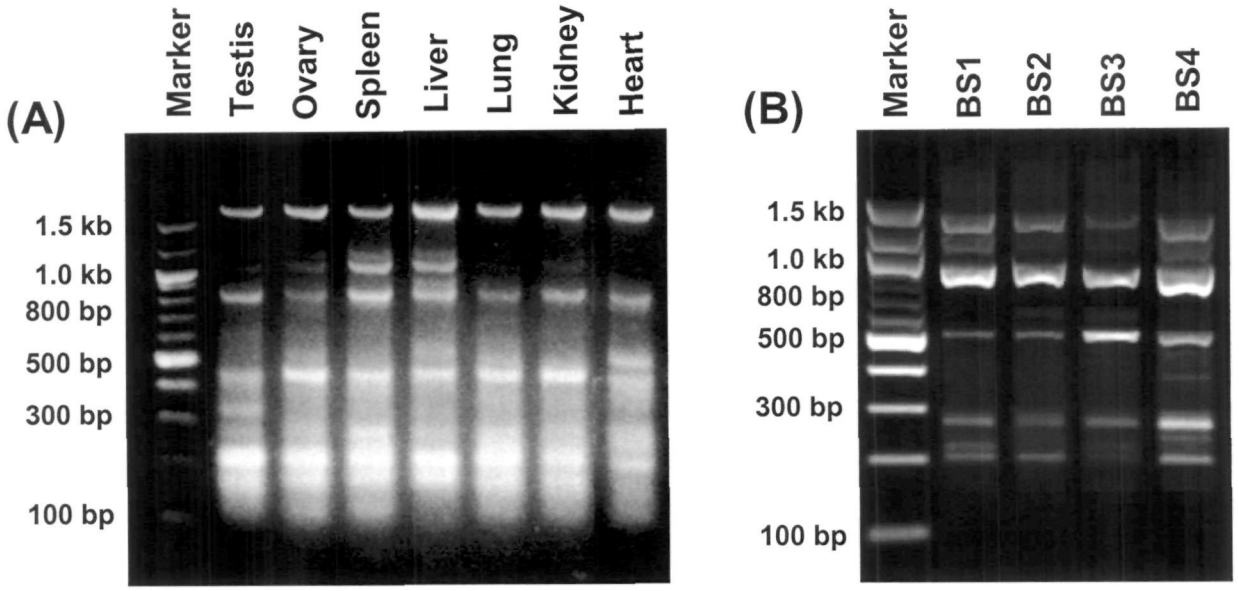


Figure 8. Microsatellite associated sequence amplification (MASA) performed using oligos based on varying lengths of GACA repeats and cDNA from different sources. The amplified transcripts ranged from 0.15 kb to 1.8 kb. MASA using GACA repeat with cDNA from different somatic and gonadal tissues is given in **(A)** and cDNA from spermatozoa from 4 animals in **(B).**

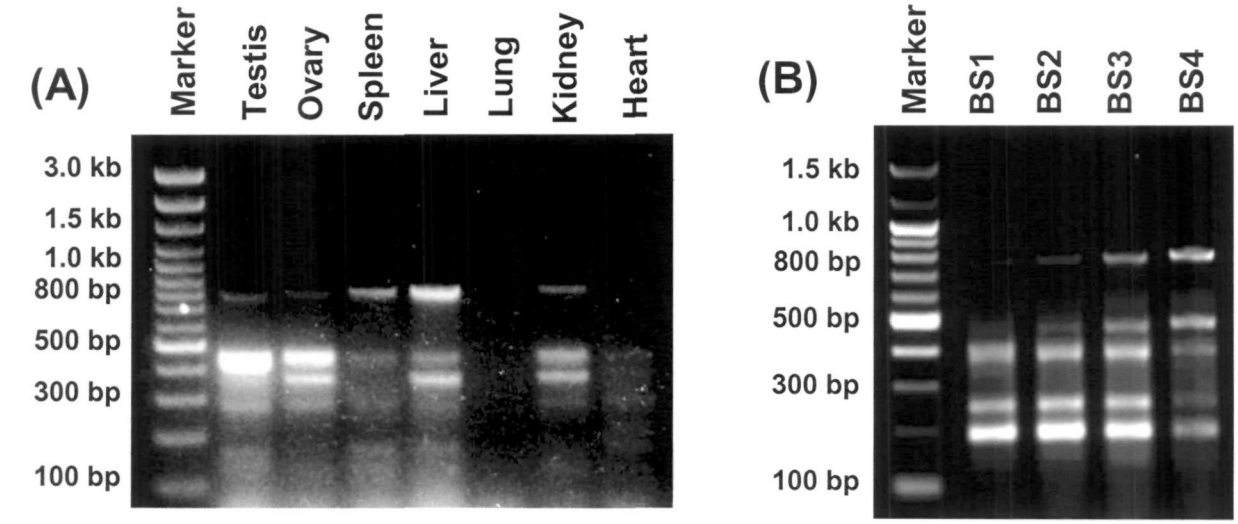


Figure 9. MASA reactions were also performed using the GATA repeats and cDNA from different somatic tissues **(A)** and the spermatozoa is shown in **(B)**. Note, the tissue and spermatozoa-specific transcript profiles generated by GACA and GATA repeats. GATA did not detect any transcripts in lung and heart.

(MASA) reactions to uncover the transcriptomes of water buffalo *Bubalus bubalis.* MASA reactions (Figure 7-9) were carried out using varying length oligos based on these repeats as primers (Table 1) and cDNA for different somatic tissues, gonads and spermatozoa as templates. Also, the MASA were performed using genomic DNA from different tissues using these repeats to compare the obtained band profile (Figure 10). All the uncovered fragments were cloned in to pGEMT-easy vector (Figure 11A) and all the recombinant clones were confirmed using Slot blot hybridization (Figure 11B), restriction digestion followed by Southern hybridizations (Figure11C-D). However, only representative figures have been shown here to show the obtained results. The possibility of interclonal variations was ruled out by restriction analysis of 6 recombinants of each fragment using different online bioinformatics softwares (http://tools.neb.com/NEBcutter2 /index.php) to predict the restriction sites (Figure 12 A-F), following the restriction digestions using combinations of enzymes (Figure 13A-H) accordingly. All the positive clones were sequenced and the sequences were submitted to the GenBank. The obtained accession numbers of all the MASA uncovered transcripts with their detailed information has been given in the tables 10-12. Database search was conducted to ascertain homology of all the sequences independently with other entries in the GenBank.

4.1.2.1 *MASA with consensus sequence of 33.15 repeat loci*

MASA using oligos based on 33.15 repeat uncovered a total of 25 amplicons representing 7 different transcripts from somatic tissues, testes and ovaries from four different animals (Figure 7A) whereas 48 amplicons comprising 12 types of transcripts were identified from spermatozoa (Figure 7C). MASA reactions were also performed using genomic DNA from different tissues to compare the profile obtained (Figure 10B). Following cloning and sequencing of all the seven tissue originated transcripts, we observed three transcripts present in all the tissues whereas one exclusively detected in each testis, spleen and liver. All the six recombinant clones for each transcript showed similar band profiles after digestion with various restriction enzymes (Figure 13A-H) indicating

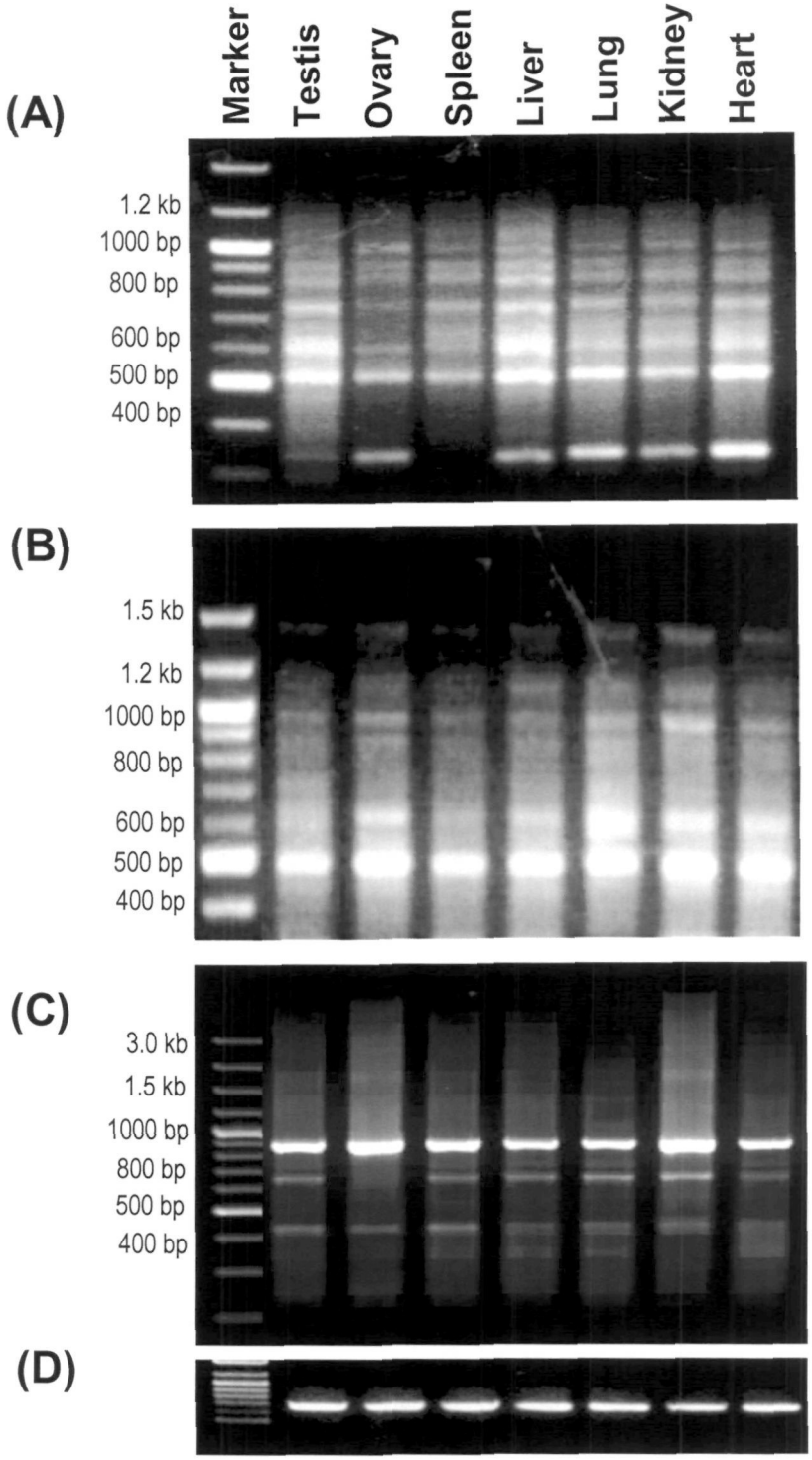


Figure 10. Minisatellite/Microsatellite associated sequence amplification (MASA) performed using oligos based on varying lengths of the 33.15, GACA and GATA repeats and genomic DNA from different sources. The amplified fragments ranged from 0.15 kb to 1.8 kb. MASA using 33.15 repeat with genomic DNA have been given in **(A),** using GACA repeat in **(B)** and using GATA repeat in **(C)**. '**D**' represents the normalization of genomic DNA samples with β-actin.

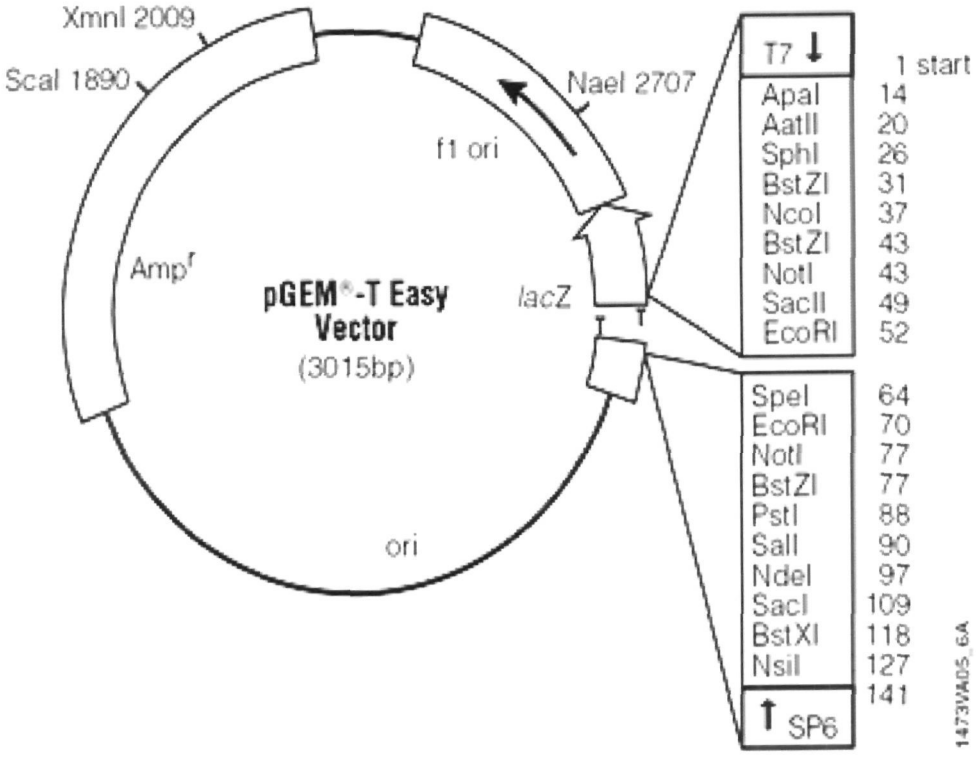


Figure 11A. Schematic map of the pGEMT- easy vector which was used for cloning all MASA uncovered fragments. Cloning site including different enzymes' sites has been also shown in the figure.

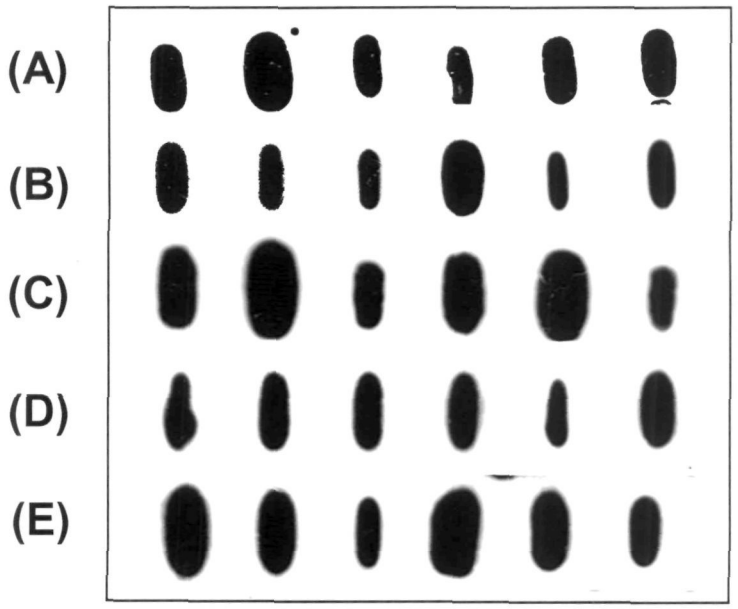


Figure 11B. Representative Slot blot hybridizations using recombinant plasmids with the genomic DNA probes **(A-E)**. The 10-12 recombinants were generated for every fragment for the slot-blot analyses.

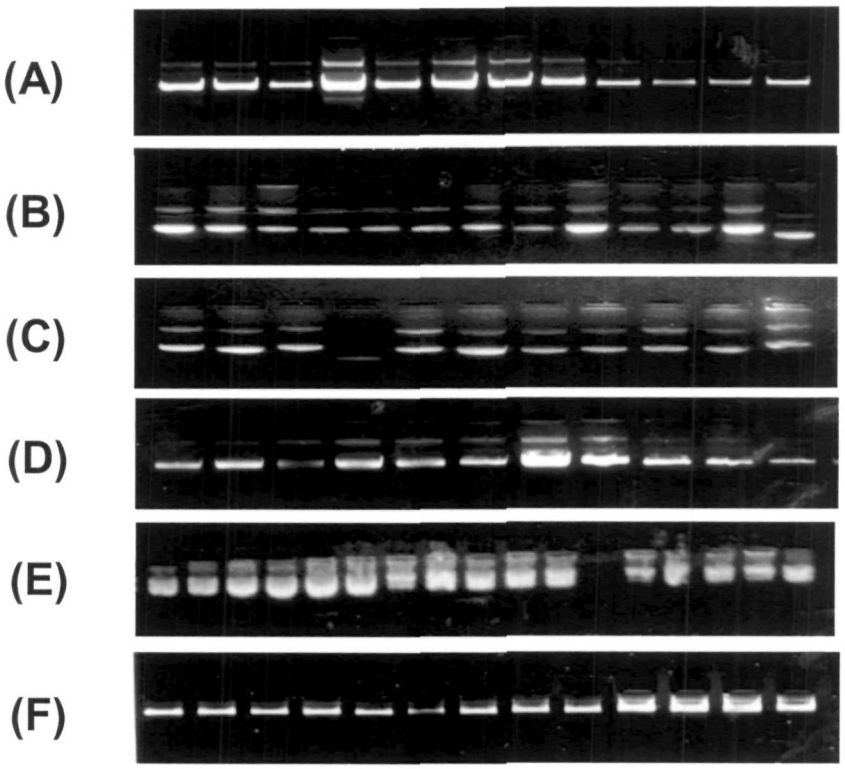


Figure 11C. Representative gel pictures showing the uncut recombinant plasmids **(A-F)** containing the genes of interest. Approximately, 10-12 plasmid DNAs were isolated by alkali-lysis method for all the fragments and checked on 1% agarose gel in 0.5X TBE.

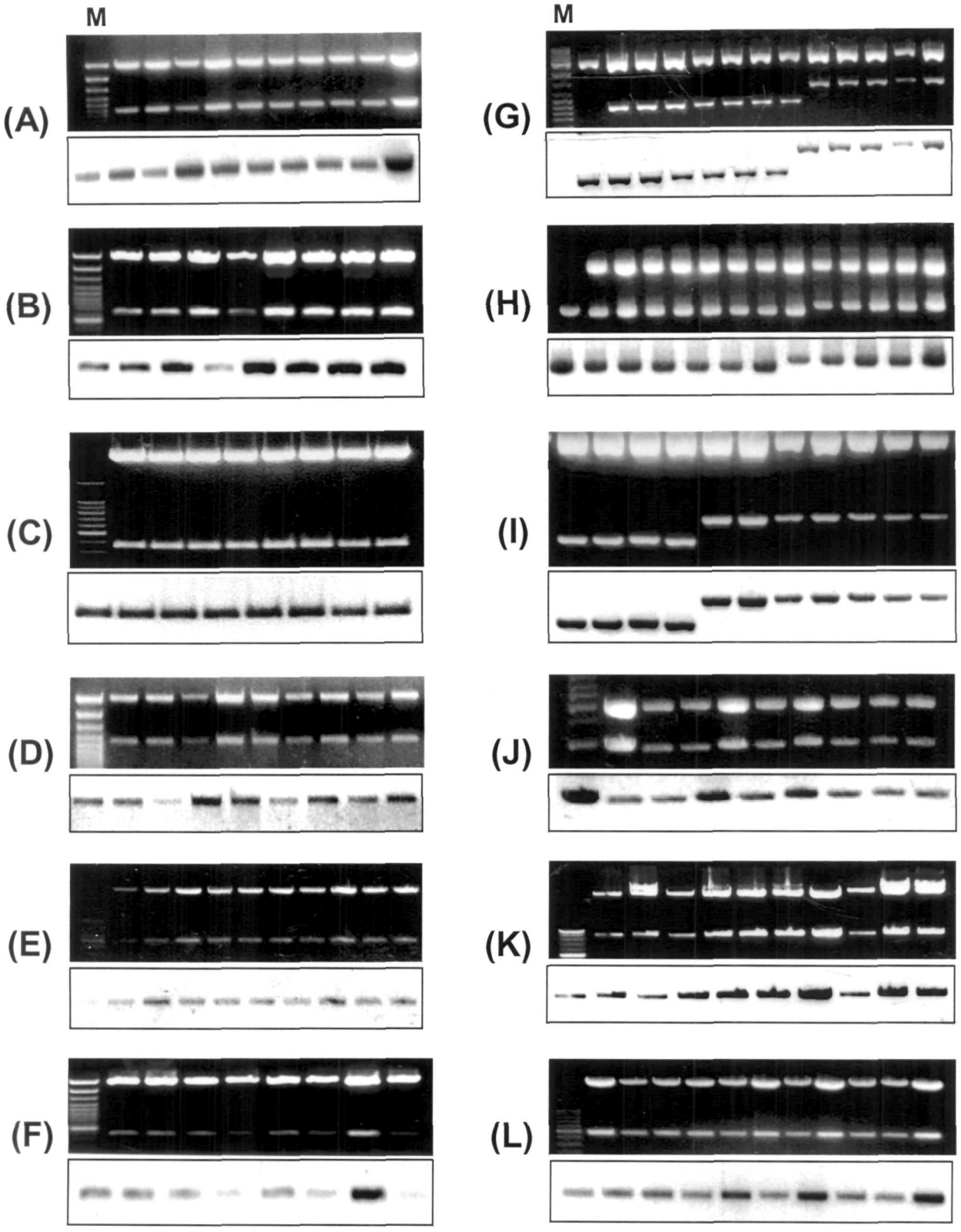
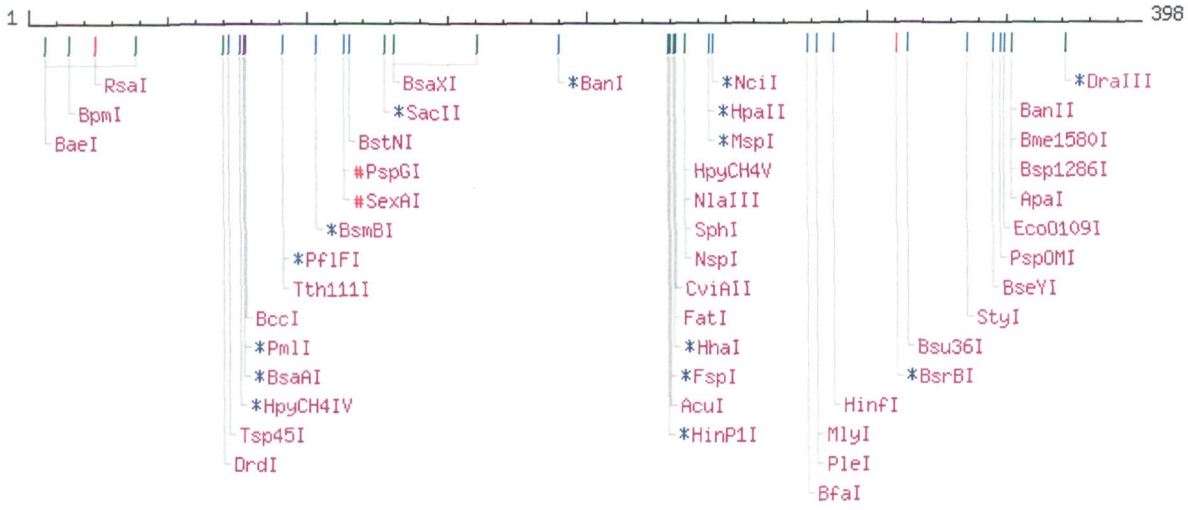
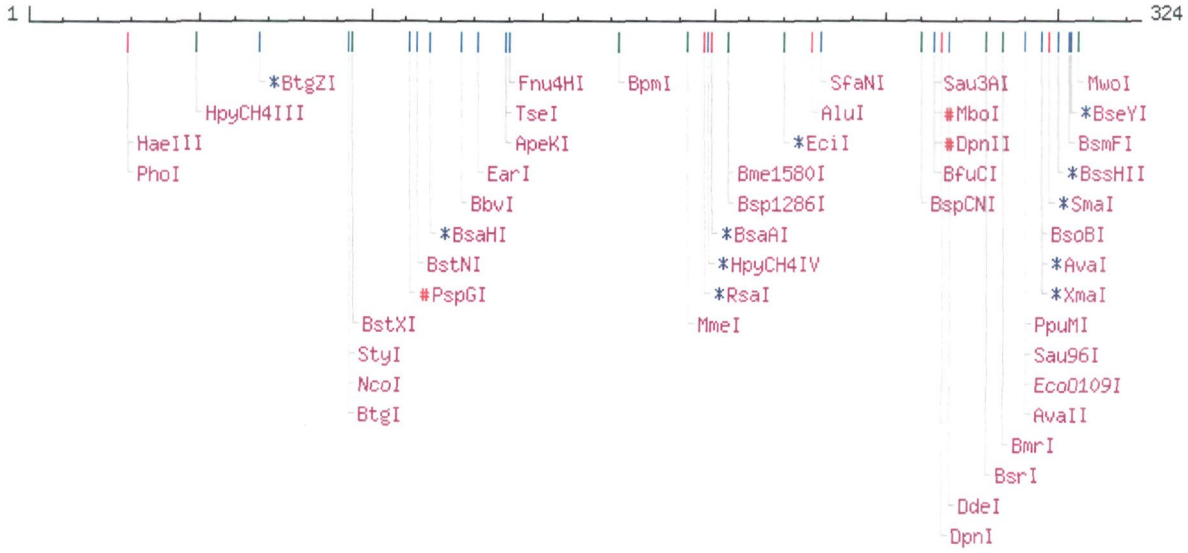


Figure 11D. Representative pictures for the restriction analyses of the recombinant clones for the MASA generated amplicons using *EcoR*1 to release the insert fragments: 850 bp **(A)**; 650 bp **(B)**; 300 bp, **(C)**; 1263 bp **(D)**; 500 bp, **(E)**; 400 bp, **(F)**; 1.1 kb and 1.8 kb **(G)**; 1.3 kb **(H)**; 350 bp and 700 bp **(I)**; 2.5kb **(J)** 1.0 kb **(K)** and 450 bp **(L)**. The autoradiograms for respective restriction digestion has also been shown below every gel photograph. However, *Rsa*I, *Hin*fI and *Hae*III enzymes were also used for these restriction analyses which showed absence of interclonal variation. The molecular size marker "M" is given in base pairs. .

(pJC3)



(pJC4)

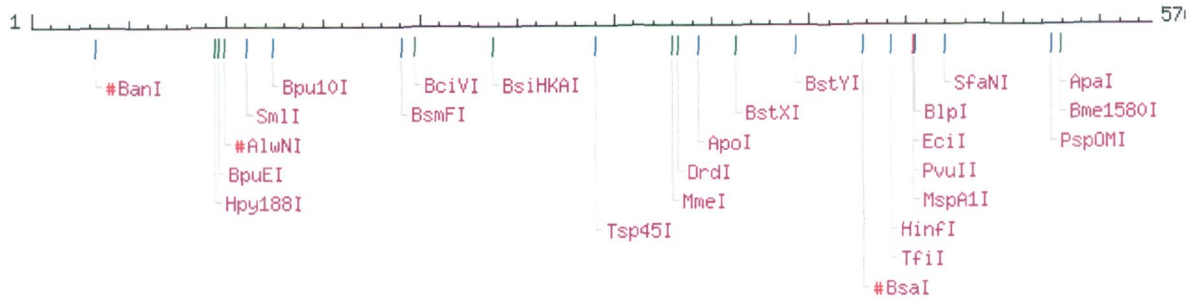


(pJC7)

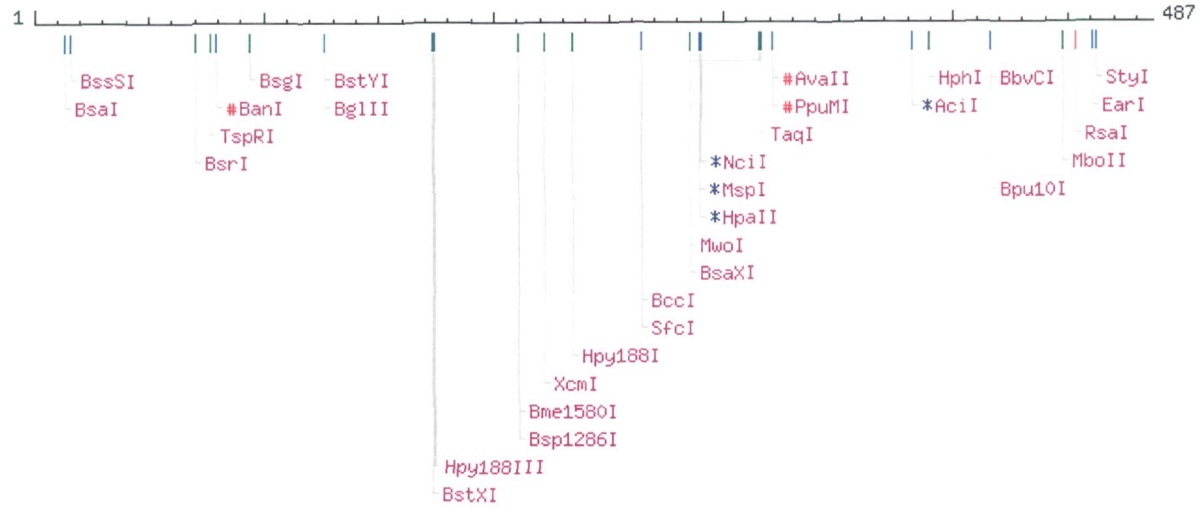
Figure 12

Contd//

(pJC5)



(pJC6)



(pJC10)

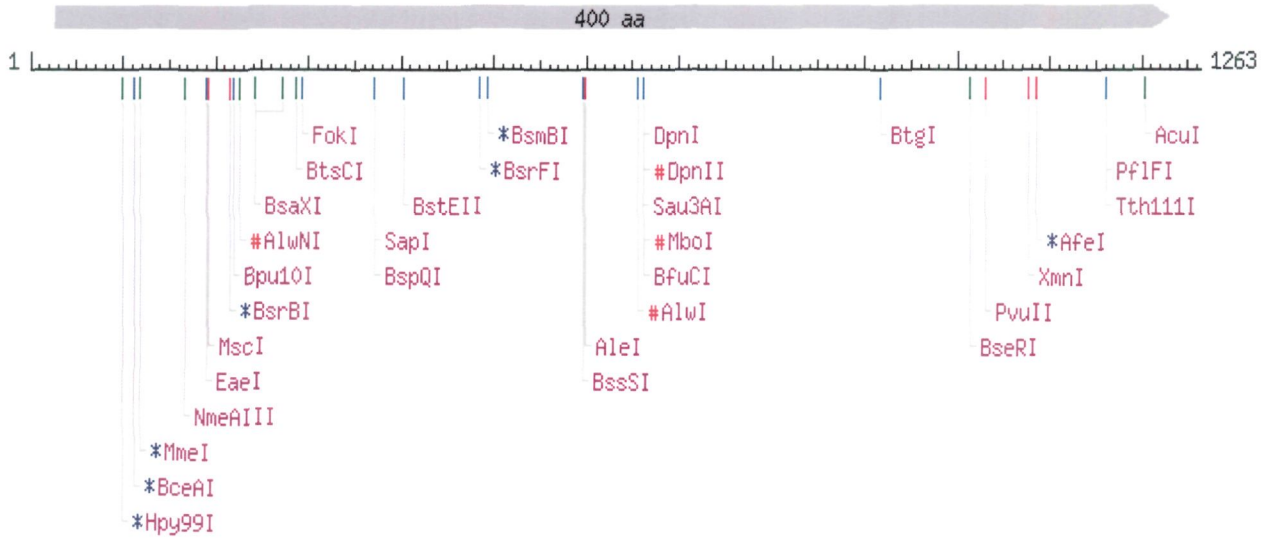


Figure 12. Restriction mapping of each fragments uncovered with the 33.15, GACA and GATA repeats was done using NEB cutter (www.tools.neb.com/NEBcutter2). Here, the representative pictures have been shown for mapping of the 33.15 uncovered fragments, and the clone IDs for respective fragments are also given on the top.

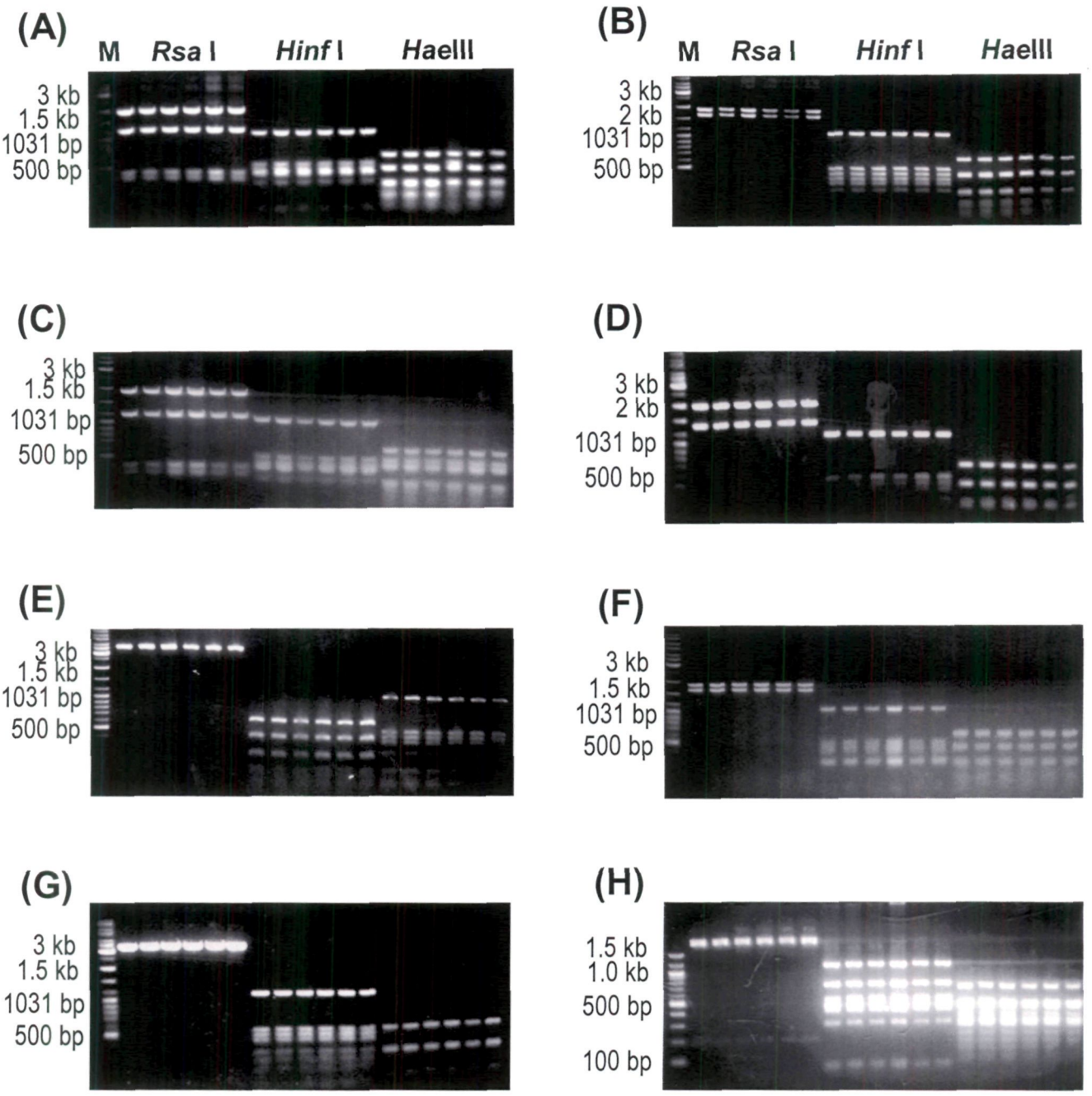


Figure 13. Restriction Analyses of the recombinant clones containing different insert fragments: 846 bp, **(A)**; 487 bp, **(B)**; 576 bp, **(C)**; 324 bp, **(D)**; 602 bp **(E)**, 579 bp **(F)**, 398 bp **(G)**, and 1263 bp **(H)** with *Rsa*I, *Hin*fI and *Hae*III enzymes showing absence of interclonal variation. The molecular size marker "M" is given in base pairs.

the absence of any inter-clonal variation. Database search using complete sequences of all these seven transcripts demonstrated their homology within and across the species (Table 10). Of these seven, three represented the similarity along with their entire length to the characterized genes for e.g. LRRN6A whereas other transcripts were homologous by partial sequences with a number of genes (for details please refer table 10).

Also, the cloning and sequencing was done individually for all the 12 transcripts uncovered from the spermatozoa. The sequence analyses and homology search for these spermatozoal transcripts showed that out of 12, 3 transcripts showed no homology, 7 were homologous to the non-characterized BAC clones originated from human, cattle or mouse. Only one was found to be similar to the Bovine steroid 21-hydroxylase gene (P-450-c21) gene and other one with *Diceros bicornis* minor clone DB5 microsatellite sequence (Table 10). Most interestingly, the comparative analysis of these transcripts with that originated from the somatic and gonadal tissues revealed that none of the gene was found to be common between the tissues and the spermatozoa. In order to ascertain the structural, functional and regulatory status of the MASA uncovered genes/gene fragments, we conducted a comprehensive database search with each one separately. The detail of these homologous genes, their accession numbers, chromosomal positions (if available), species, position of our cDNA sequences and their possible functions are given in the table 10.

4.1.2.2 *Simple repeat of GACA uncovered more number of transcripts in spermatozoa than in somatic/gonadal tissues*

MASA using GACA repeat identified a total of 57 amplicons, each from four animals representing 14 different types of transcripts from somatic tissues and gonads (Figure 8A and Table 11). Also, 104 amplicons were detected in spermatozoa representing 26 types of transcripts (Figure 8B and Table 11). Upon subsequent cloning and characterization, we observed that of the 14 tissue-originated transcripts, only 5 were common to all the tissues while remaining ones showed

Table 10: Detailed analyses of the mRNA transcripts tagged with consensus of 33.15 repeat originated from buffalo *Bubalus bubalis* with genes across the species#

(A) mRNA transcripts uncovered from different tissues								
Clone ID	**Accession no.**	**Origin/Size (In bp)**	**Homology Status**	**Accession no. of the homologues**	**Gene length (in bp)**	**Chromo -somal position**	**Position of uncovered transcripts**	**% Homo - logy**
pJC4 pJC27 pJC28	AY762112 DQ017890 DQ017889	Testis/324 Spleen/324 Lung/324	1. *Homo sapiens* leucine rich repeat neuronal 6A(LRRN6A), mRNA	NM_032808	2932	15q24.3	1657-1946	96%
			2. *Pan troglodytes* similar to leucine-rich repeat neuronal 6A (LOC453764), mRNA	XM_510686	2811	15	2553-2807	96%
			3. *Canis familiaris* similar to leucine-rich repeat neuronal 6A (LOC487660), mRNA	XM_544785	2196	30	1938-2196	96%
			4. *Mus musculus* leucine rich repeat neuronal 6A, mRNA	BC052384	3321	9 A3	2057-2320	92%
			5. *Gallus gallus* similar to leucine-rich repeat neuronal 6A, mRNA	XM_413730	2083	10	1708-1946	89%
			6. *Rattus norvegicus* leucine rich repeat neuronal 6A	XM_236268	3656	8	2380-2637	93%
			7. *Macaca fascicularis* brain cDNA clone	AB046639	2621	----	2030-2314	95%
			8. *Mus musculus* male testis cDNA	AK027262	2290	----	1680-1893	92%
pJC6 pJC16 pJC17 pJC18 pJC19 pJC22 pJC23	AY762115 DQ011908 DQ011909 DQ011910 DQ011911 DQ011914 DQ011915	Spleen/487 Kidney/487 Liver/487 Heart/487 Lung/487 Testis/487 Ovary/487	1.*Homo sapiens* micro-satellite-associated sequence amplification product	AF134482	513	Y	17-503	99%
			2. *Homo sapiens* chromosome Y palindromes P1, P2, P3 and inverted repeat IR2 (P1-P2-P3-IR2@) on chromosome Y	NG_004755	4538292	Yq11.23	484366-484844	100%
			3. *Homo sapiens* BAC clone RP11-47785 from Y	AC007320	187894	Y	32965-33443	100%
			4. *Pan troglodytes* BAC clone RP43-31J1 from Y	AC146245	189042	Y	121813-122286	86%

pJC5	AY762114	Spleen/576	1. *Bos taurus* pancreatic anionic trypsinogen (TRYP8) gene	AF453325	151482	----	53697-53881	89%
			2. *Bos taurus* Major histocompatibility complex, class II, DM beta-chain	XM_58348	1590	----	1177-1357	88%
			3. *Bos taurus* similar to unc-13 homolog D (LOC511333), mRNA	XM_588648	776	----	----	88%
			4. *Bos taurus* T cell receptor gamma (TCRG), gene	AY644517	125381	----	----	87%
			5. *Bos taurus* Tcell receptor alpha (TCRA) gene, J fragments and C region	AY227782	97500	----	66122-66301 56705-56902	87%
			6. *Bos taurus* argininosuccinate synthetase (ASS) gene	AY055202	1367	11	----	87%
			7. *Ovis aries* beta- globin gene	OABBGLOB	3249	----	----	86%
			9. *Bos taurus* similar to epithelial chloride channel protein (LOC507643), partial mRNA	XM_584294	3238	----	----	83%
pJC8 pJC20 pJC21 pJC24 pJC25 pJC26	AY920927 DQ011912 DQ011913 DQ017885 DQ017886 DQ017887	Testis/602 Kidney/602 Heart/602 Ovary/602 Liver/602 Lung/602	1. *Bos taurus* T cell receptor gamma (TCRG) gene, partial sequence	AY644517	125381	----	96417-96530 30298-30429	85%
			2. *Bos taurus* pancreatic anionic trypsinogen (TRYP8) gene, complete cds; germline T-cell receptor beta DJC region genes	AF453325	151482	----	10518-10632 119738-119844	85%
			3. *Bos taurus* epidermal growth factor precursor (EGF) gene, exons 19	AY192564	7387	----	6174-6211 2362-2409	100%
			4. *Bos taurus* X-inactivation center region, Jpx and Xist genes	AJ421481	233345	----	1382-1433	94%
			5. *Bos taurus* X (inactive)-specific transcript (XIST) on chromosome X	AJ421481	233345	----	1382-1433	95%
			6. *Bos taurus* solute carrier family 35 member 3 (slc35a3) gene	AY160683	22404	----	----	92%
			7. *Bos taurus* bone morphogenetic protein receptor IB gene, exons 8 and 9	AY242067	1253	6 "44.2 cM"	855-955	86%
			8. *Ovis aries* microsatellite sequence INRA054	AF259768	824	----	437-547	84%
pJC7 pJC9	AY847460 AY899285	Testis/846 Spleen/847	1. *Bos taurus* adenylate kinase isozyme 2 (EC 2.7.4.3) gene, exon 6 and 7	D90069	7816	---	4627-4821	91%

pJC11 pJC12 pJC13 pJC14 pJC15	AY945208 AY997303 AY997304 DQ011906 DQ011907	Ovary/846 Kidney/847 Liver/847 Lung/847 Heart/847	2. *Bos taurus* Mx1 gene for GTP-binding protein	AB060171	2079	18q26	1004-1200	89%
			3. *Bos taurus* tnni3 gene for cardiac troponin I, exons 1-8	AJ842179	6339	18q26	3928-4087	91%
			4. *Bos taurus* bcnt, h-type bcnt genes	AB081095	92594	18	7942-8134	89%
			5. *Bos taurus* partial ed1 gene for ectodysplasin A, exons 4-9	AJ278907	37331	Xq22-q24	10503-10697	89%
			6. *Bos taurus* pancreatic anionic trypsinogen (TRYP8) gene, complete cds; germline T-cell receptor beta DJC region genes, partial cds	AF453325	151482	----	138541- 138734	89%
			7. *Bos taurus* calpastatin (CAST) gene, exon 1u	AH014526S04	6712	----	----	88%
			8. *Bos taurus* lysozyme gene	BOVLYSOZMA	12222	----	----	88%
			9. *Capra hircus* sex-specific gonadal PISRT1 mRNA	AF404302	48420	1	----	85%
pJC10	AY947405	Liver/1263	1. *Homo sapiens* SPARC related modular calcium binding 1, mRNA	BC011548	1960	14q24.2	345-1591	93%
			2. *Pan troglodytes* similar to secreted modular calcium-binding protein 1 (LOC453002), mRNA	XM_510036	2454	14	860-2103	92%
			3. *Rattus norvegicus* SPARC-related modular calcium binding protein 1 (Smoc1), mRNA	NM_00100283 5 XM_243307	1359	6q24	83-1283	89%
			4. *Mus musculus* SPARC related modular calcium binding 1 (Smoc1), mRNA	BC031804	3460	12 C3	314-1589	89%
			5. *Bos taurus* similar to Secreted modular calcium-binding protein 1 (LOC532833), partial mRNA	XM_612029	2549	----	1-674	98%
			6. *Canis familiaris* similar to Secreted modular calcium-binding protein 1 (LOC480373), mRNA	XM_537495	4031	8	3014-3690	92%

(B) mRNA transcripts detected in the Spermatozoa

Clone ID	Accession no.	Size(bp)	Homology Status	Accession no. of the homologues	Gene length	Chromo-somal position	Position of uncovered mRNA trans.	% Homo-logy
pJSC39	EU348479	276 bp	1. *Homo sapiens* cDNA FLJ30354 fis, clone BRACE2007682	AK054916	1698	----	590-699	76%
			2. *Homo sapiens* chromosome 19 clone CTD-2292J18	AC090427	109233	19	72209-72318	76%
pJSC40	EU348480	276 bp	1. Bovine steroid 21-hydroxylase gene (P-450-c21) gene	BOVP45C21	6601	---	6323-6583	94%
			2. *Ovis aries* cytochrome P-450 steroid 21 hydroxylase (P-450-c21) gene	EF382834	3454	20	3009-3272	94%
			3. Sheep cytochrome protein (P450C21C) mRNA	SHPCYTOC	2020	--	1862-2018	96%
pJSC41	EU348481	270 bp	1. *Homo sapiens* BAC clone RP11-498O22 from 2	AC012506	197303	2	172668-172905	69%
pJSC42	EU348482	267 bp	No homology					
pJSC43	EU348483	297 bp	No homology					
pJSC44	EU348484	522 bp	1. Human DNA sequence from clone RP1-230L10 on chromosome 6 Contains part of a novel gene	AL137005	103679	6	54567-54689	76%
pJSC45	EU348485	517bp	1. *Bos taurus* BAC CH240-319N17 (Children's Hospital Oakland Research Institute Bovine BAC Library (male))	AC150512	167026	--	164216-164336 135400-135520 148267-148376 61276-61394 100657-100775	96% 95% 96% 93% 92%
			2. *Homo sapiens* genomic DNA, chromosome 11 clone:RP11-687M24	AP001007	203840	11	185100-185294	83%
			3. Bovine alpha s2 casein type A protein (CASAS2) gene, exons 1-18	BOVCASAS2X	21146	---	17455-14574 9535-9624	94% 87%

			4. *Bos taurus* clone p68.3 MHC class I antigen gene, 5' UTR and partial cds	AF396777	1169	---	548-668	95%
pJSC46	EU348486	537 bp	1. *Rhinolophus ferrumequinum* clone VMRC7-44G7	AC149031	192138	---	44739-45277	68%
			2. *Felis catus* clone RP86-177C24	AC087421	141881	---	4255-4688	70%
			3. *Canis familiaris* MHC class II region on chromosome CFA12, partial sequence, BAC clone 181g17, containing DLA-DMB, DLA-DMA and BRD2 genes, GL004-like protein pseudogene, DLA-DOA gene, RPL26 and RPL11 pseudogenes, DLA-DPB1 pseudogenes, DLA-DPA1 pseudogene, UPF0224 family pseudogene and partial col11a2 gene	AJ630365	155570	CFA12q	92839- 93370	68%
pJSC47	EU348487	387 bp	1. *Homo sapiens* chromosome 15 clone RP11-358M11 map 15q21.3	AC016525	256019	15q21.3	119191-119420	75%
pJSC48	EU348488	398 bp	1. *Bos taurus* BAC CH240-96I24 (Children's Hospital Oakland Research Institute Bovine BAC Library (male)	AC213711	186273	---	149863-150033 117725-117896 29045-29213 177031-177200	96% 94% 91% 91%
			2. *Bos taurus* Fanconi anemia, complementation group M, mRNA	BC149286	1448	---	1228-1398	96%
			3. *Bos taurus* F-box protein 17, mRNA	BC149053	4223	---	3224-3404	95%
			4. *Bos taurus* BTA13 scaffold140080_30362 genomic sequence contig containing highly polymorphic single nucleotide sites	EF034081	19543	13	9387-9557 14352-14522	95% 95%
			5. *Bos taurus* isolate BSX MHC class I antigen gene	AF396751	4604	---	2105-2275	95%
pJSC49	EU348489	370	1. *Bos taurus* BAC CH240-10G15 (Children's Hospital Oakland Research Institute Bovine BAC Library (male)	AC149774	186916	---	129128-129350	85%
			2. *B.taurus* DNA for SINE sequence Bov-2	X64125	560	---	182-400	84%

			3. *O.aries* KII-9 gene for hair type II keratin intermediate filament	X62509	7371	---	1823-2030	83%
			4. *Bos taurus* similar to Myotubularin-related protein 2, mRNA	BC148076	5057	---	4447-4665	82%
pJSC50	EU348490	375	No homology					

The transcripts uncovered from somatic and gonadal tissues are given in **(A)** whereas spermatozoal transcripts in **(B)**. All of the 33.15-tagged transcripts were submitted to the GenBank and the accession numbers were obtained for each transcript. The analysis carried out for their homologues, size and chromosomal positions is also given in the table.

Table 11: Detailed analyses for the MASA identified somatic and spermatozoal transcripts tagged with GACA repeat motif from water buffalo *Bubalus bubalis#*

(A) mRNA Transcripts uncovered from different tissues								
Clone ID	**Accession no.**	**Tissue origin /Size(bp)**	**Homology Status**	**Accession no. of the homologues**	**Gene length**	**Chromo-somal position**	**Position of uncovered mRNA trans.**	**% Homo-logy**
pJC29	DQ289479	Brain/1769	1. *Bos taurus* target 1 genomic scaffold	DP000008	2072671	-	109-395	90%
pJC30	DQ289480	Heart/1768	2. *Bos taurus* lactoferrin (Lf) gene, 5' flanking region exons 1, 2	AY319306	8212	22	123-385	90%
pJC31	DQ289481	Liver/1768						
pJC32	DQ289482	Lung/1812	3. *Bos taurus* T-cell receptor gamma cluster 2 (TCRG2) gene	AY644518	188109	-	109-386	89%
pJC33	DQ289483	Ovary/1767						
pJC34	DQ289484	Spleen/1772	4. *Bos taurus* prion preproprotein (PRNP) and prion-like protein doppel preproprotein gene	AY944236	207929	-	131-395	90%
pJC43	DQ494486	Testis/1767						
pJC55	NS	Kidney/1812	5. *Bos taurus* glutamate-cysteine ligase catalytic subunit (GCLC)	AY957499	447010	-	109-356	91%
pJC44	DQ534902	Kidney/1303	1. Pig DNA sequence from clone CH242-277I8	CR956634	206278	17	104-277	86%
pJC45	DQ534903	Liver/1303	2. Human DNA sequence from clone RP5-1009H6 on chromosome 20 Contains the 3' end of the NFATC2 gene for cytoplasmic calcineurin-dependent (2) nuclear factor of activated T-cells	HS1009H6	89163	20	158-245 627-774	90%
pJC56	NS	Ovary/1303						
pJC57	NS	Spleen/1303						
pJC58	NS	Testis/1303						
pJC35	DQ304116	Heart/1080	1. Human DNA sequence from clone RP11-148E14 on chromosome 10 Contains part of the BTRC gene for beta-transducin repeat	AL627144	36454	10	281-884	94%
pJC36	DQ304117	Liver/1080						
pJC37	DQ304118	Lung/1080						
pJC38	DQ494481	Ovary/1080	2. *Mus musculus* BAC clone RP23-408K9 from chromosome 19	AC140332	206515	19	282-884	90%
pJC39	DQ494482	Spleen/1080						
pJC41	DQ494484	Kidney/1080						
pJC59		Testis/1080						
pJC40	DQ494483	Testis/1043	1. *Bos taurus* prion protein (PRNP) and prion-like protein doppel (PRND) genes, PRNT gene, exons 1 and 2; and putative protein gene	DQ205538	104027	13q17	333-857	89%

			2. *Ovis aries* prion protein gene	U67922	31412	-	333-857	88%
			3. *Odocoileus hemionus* prion protein (prnp) gene	AY330343	65476	-	333-857	87%
pJC42	DQ494485	Kidney/1067	1. *Bos taurus* similar to ring finger protein 149 (LOC506267)	XM_582694	4148	-	398-597	97%
			2. *Canis familiaris* similar to ring finger protein 149	XM_538454	1152	10	403-446	93%
pJC46 pJC60 pJC61 pJC62 pJC63 pJC64 pJC65	DQ534904 NS NS NS NS NS NS	Liver/848 Spleen/850 Heart/848 Testis/848 Kidney/848 Ovary/848 Lung/858	1. *B.taurus* mRNA HBGF-1 for acidic fibroblast growth factor (5'end)	X66446	412	-	131-441	97%
			2. *Bos taurus* fibroblast growth factor, acidic (FGF1), mRNA	NM_174055	4005	7	131-374	97%
			3. *Bubalus bubalis* clone BBMS119 microsatellite sequence	AY779568	452	-	68-285	100%
			4. *Homo sapiens* gene for acidic fibroblast growth factor	Z14150	1185	-	256-842	86%
pJC54	DQ834345	Testis/725	1. *Bos taurus* target 1 genomic scaffold	DP000008	2072671	-	139-261	90%
			2. *Bos taurus* bone morphogenetic protein receptor IB gene, exons 8 and 9	AY242067	1253	6	139-261	86%
			3. *Bos taurus* testis expressed sequence 10, mRNA	BC112672	2828	-	174-253	91%
pJC49 pJC66 pJC67 pJC68 pJC69 pJC70	DQ534907 NS NS NS NS NS	Ovary/635 Kidney/635 Heart/635 Liver/635 Testis/647 Spleen/635	1. Human DNA sequence from clone RP4-752I6 on chromosome 1 Contains the 5' end of the WASF2 gene for WAS protein family	BX293535	71971	1	445-485 555-635	91%
			2. Mouse DNA sequence from clone RP23-125F21 on chromosome 4	AL627184	152069	4	555-635	90%
pJC50 pJC71	DQ534908 NS	Spleen/612 Ovary/612	1. *Bos taurus* similar to ankyrin repeat domain 26	XM_580719	1470	21	119-368	86%
*pJC48 pJC72	DQ534906 NS	Testis/523 Ovary/523	1. *Bos taurus* similar to ankyrin repeat domain 26	XM_580719	1470	21	156-405	86%
pJC47 pJC73	DQ534905 NS	Brain/455 Heart/455	1. *Bubalus bubalis* clone 2 minisatellite sequence	AY230133	419	-	43-437	100%

pJC74 pJC75 pJC76 pJC77 pJC78 pJC79	NS NS NS NS NS NS	Kidney/455 Ovary/455 Spleen/455 Lung/455 Testis/455 Liver/455	2. *Homo sapiens* 12 PAC RPCI1-53O8	AC005344	153836	12	125-251	86%
pJC53	DQ834344	Heart/412	1. *Bos taurus* DNA for SINE sequence Bov-tA	X64124	197	-	52-224	89%
			2. *Bos taurus* ABCG2 gene, PKD2 gene and SPP1 gene, clone RPCI42_5K14	AJ871176	171712	6	52-233	86%
			3. *Bos taurus* similar to ataxin-1 ubiquitin-like interacting protein, transcript variant 6	XM_882781	3406	3	54-116	87%
			4. *Bos taurus* BTA01 genomic sequence contig containing highly polymorphic single nucleotide sites	DQ404150	5376	1	84-245`	87%
pJC51 pJC80 pJC81 pJC82 pJC83 pJC84 pJC85	DQ534909 NS NS NS NS NS NS	Testis/209 Liver/209 Lung/209 Ovary/209 Spleen/209 Kidney/209 Heart/209	1. *Mus musculus* chromosome 1, clone RP23-474A1	AC163217	184175	1	186-209	100%
			2. *Mus musculus* BAC clone RP24-114C10 from chromosome 13	AC165149	191162	13	188-209	100%
*pJC52	DQ534910	Testis/217	1. *Bos taurus* similar to Ubiquitin-associated protein 1, transcript variant 2	XM_865289	4601	8	7-207	99%
			2. *Canis familiaris* similar to Ubiquitin-associated protein 1, transcript variant 1	XM_531976	2660	11	37-207	94%
			3. *Macaca mulatta* ubiquitin associated protein 1 (UBAP1),	XM_001089450	4100	15	9-207	90%
			4. *Homo sapiens* ubiquitin associated protein 1 (UBAP1),	NM_016525	2752	9p13.3	9-207	90%

(B) Transcripts identified in the spermatozoa

Clone ID	Accession no.	Size(bp)	Homology Status	Accession no. of the	Gene length	Chromo-somal	Position of uncovered	% Homo-

				homologues		position	mRNA trans.	logy
pJSC1	DQ789045	1313	▪ Same as pJC44-45 and pJC56-58					
pJSC2	DQ789046	857	▪ Same as pJC46 and pJC60-61					
pJSC3	DQ789047	807	1. *Bubalus bubalis* minisatellite associated amplified segment	AY212951	757	-	16-792	96%
			2. *Bos taurus* similar to non-POU domain containing, octamer-binding	BC105532	2580	-	558-737	90%
pJSC4	DQ789048	789	1. Hippopotamus amphibius DNA, SINE-containing sequence	AB007204	311	-	582-611	100%
			2. *Bos taurus* BTA29 11629 genomic sequence contig containing highly polymorphic single nucleotide sites	DQ404153	18838	29	659-686	100%
			3. *Globicephala macrorhynchus* DNA, CHR-2 SINE FL type sequence	AB071578	321	-	659-742	88%
pJSC5	DQ789049	844	1. *Bos taurus* similar to zinc finger, DHHC domain	XM_869440	1676	-	217-414	91%
			2. *Canis familiaris* similar to zinc finger, DHHC domain	XM_846705	1470	-	290-397	83%
pJSC6	DQ834346	797	1. *Homo sapiens* BAC clone RP11-703G6 from 4	AC074349	176467	4	95-401	85%
pJSC7	DQ834347	840	▪ Same as pJC48 and pJC50			-		
pJSC8	DQ845141	635	▪ Same as pJC49 and pJC66-70			-		
pJSC9	DQ845142	507	1. *Bos taurus* prion preproprotein (PRNP) and prion-like protein doppel preproprotein (PRND	AY944236	207929	-	52-339	88%
			2. *Bos taurus* T cell receptor gamma cluster 2 (TCRG2) gene	AY644518	188109	-	52-339	87%
			3. *Capra hircus* sex-specific gonadal PISRT1 mRNA	AF404302	48420	1q43	52-337	87%
pJSC10	DQ845143	516	1. *Bos taurus* similar to Disabled homolog 2	BC111684	805	-	272-443	97%
			2. *Homo sapiens* disabled-2 gene	AF218839S1	2196	5p12-p13	356-507	91%
			3. *Pan troglodytes* similar to disabled 2 p93	XM_517792	5113	5	356-435	91%
pJSC11	DQ845144	523	▪ Same as pJC48, pJC50 and pJSC6			-		
pJSC12	DQ845145	532	1. Human DNA sequence from clone RP11-790G19 on chromosome 10 Contains the 5' end of the gene for transmembrane receptor Unc5H2, the 3'end of a novel gene and two CpG islands	AL359832	195130	10	394-431	97%

pJSC13	DQ845146	531	1. *Mus musculus* chromosome 15, clone RP24-236A19	AC158973	187091	15	46-327	83%
			2. *Homo sapiens* chromosome 8, clone RP11-1077K19	AC104247	118230	8	133-377	84%
pJSC27		522	▪ Same as pJC48, 50, 71 & 72			-		
pJSC14	DQ904036	455	▪ Same as pJC47 and pJC73-79			-		
pJSC15	DQ904037	392	1. *Mus musculus* BAC clone RP23-136L14 from chromosome 16	AC166171	199601	16	362-398	100%
pJSC16	DQ904038	387	1. *B.taurus* micosatellite DNA, clone BOV1.1.2	Y07736	826	-	160-335	89%
			2. *Bos taurus* BAC CH240-275I24 (Children's Hospital Oakland Research Institute Bovine BAC Library (male)	AC150707	153353	-	120-261	90%
pJSC17	DQ904039	354	1. *Bos taurus* similar to Potassium voltage-gated channel subfamily C member 4 (Voltage-gated potassium channel subunit Kv3.4) (Raw3)	XM_613047	2561	3	33-346	97%
pJSC18	DQ913640	267	1. Zebrafish DNA sequence from clone CH211-222O4 in linkage group 3	BX004760	190220	-	2-28	96%
pJSC19	DQ913641	277	1. *Mus musculus* BAC clone RP23-111N9 from chromosome 7	AC147502	202934	7	165-191	96%
pJSC20	DQ913642	291	1. *Bos taurus* partial ed1 gene for Ectodysplasin 1	BTA300468	9596	Xq22-q24	97-220	91%
			2. *Bos taurus* HIV-1 Tat interactive protein 2 HTATIP2	BC104577	1645	-	97-218	90%
			3. *Bos taurus* similar to C4b-binding protein alpha chain precursor (Proline-rich protein) (PRP)	XM_583188	2960	-	97-216	90%
pJSC21	DQ913643	301	▪ Same as pJC48, pJC50, pJSC6 and pJSc11			-		
pJSC22	DQ913644	273	1. *Ovis aries* 5' flanking region of the Jaagsiekte Sheep Retrovirus integration site	AY322397	466	-	91-203	89%
			2. *Bos taurus* similar to NipSnap1 protein	XM_866639	2458	17	100-203	90%
			3. *Bos taurus* lysozyme (LZ) gene	U25810	12039	5q23	118-205	91%
pJSC23	DQ913645	274	1. Human DNA sequence from clone RP11-541N10 on chromosome 10 Contains the 5' end of the SH3MD1 gene for SH3 multiple domains 1, a	AL133355	190882	10	103-254	89%

			novel gene and two CpG islands					
pJC24	DQ913646	269	NA			-		
pJSC25	DQ916743	229	1. *Mus musculus* BAC clone RP23-476B3 from chromosome 7	AC121827	183470	7	1-26	100%
pJSC26	DQ916744	209	▪ Same as pJC51, pJC80-85			-		

The transcripts uncovered from somatic and gonadal tissues are given in **(A)** whereas spermatozoal transcripts in **(B)**. All of the GACA-tagged transcripts were submitted to the GenBank and the accession numbers were obtained for each transcript. The analysis carried out for their homologues, size and chromosomal positions is also given. Blast search showed homology of these transcripts with several genes/gene fragments across the species. Notably, only few of them represented by '*' had homology along the length while others showed partial homology.

tissue-specific profiles. Of these tissue-specific transcripts, 3 were exclusive to testis, 1 each for kidney and heart, 1 common for testis and ovary, and 9 absent in lung (Table 11).

Following this, we also characterized 26 transcripts detected in the spermatozoa, and found that only 6 were shared with somatic tissues whereas remaining 20 were exclusive to the spermatozoal RNA pool (Table 11). Further, approximately 80% of the somatic tissues and 50% of the spermatozoa originated transcripts showed significant homologies (greater than 85%) with several coding genes across the species or uncharacterized BAC clones originated from cattle or human. Remaining fragments showed non-substantial or no homology with the genes present in the databank. Moreover, amongst the transcripts showing homology, only two fragments showed similarity along their entire length (GenBank Accession no. DQ534910 and DQ534906) to the database representatives, whereas remaining ones were homologous either to 5'/3' or intervening sequences of the characterized genes. Interestingly, >80% of the homologous genes were found to be involved in either signal transduction or cell-cell interaction pathways whereas remaining 20% were implicated with several diseases. Details of the GACA-tagged transcripts with their accession numbers and homologous genes are given in the table 11.

4.1.2.3 *Transcripts identified by GATA repeat in different tissues and spermatozoa*

In contrast to GACA and 33.15, GATA repeat uncovered few transcripts but with well-defined tissue-specific profiles (Figure 9A). Briefly, a total of 24 amplicons each from 4 different animals constituting 10 types of transcripts were identified from gonads and other somatic tissues barring lung and heart which were conspicuously devoid of any amplicon (Table 12). These transcripts further showed tissue-specificities upon cloning and sequencing such that 6 were exclusive to testis, while remaining 4 were common to other tissues. Also, we uncovered 40 amplicons identifying 10 types of transcripts from spermatozoa (Figure 9B) which upon cloning and characterization were found to be identical to the

Table 12: Detailed analyses for the MASA uncovered somatic and spermatozoal transcripts tagged with the GATA repeat motifs from water buffalo *Bubalus bubalis*#

(A) Identified from somatic tissues and gonads				(B) Identified from spermatozoa			
S.No.	**Clone ID**	**Accession numbers**	**Origin/Size (in bp)**	**S.No.**	**Clone ID**	**Accession numbers**	**Size (in bp)**
1.	pJC86	EF051520	Kidney/807	1.	pJSC28	EF050082	808
2.	pJC95	NS	Testis/807	2.	pJSC31	EF051516	425
3.	pJC94	NS	Ovary/807	3.	pJSC30	EF050084	414
4.	pJC93	NS	Spleen/821	4.	pJSC32	EF051517	417
5.	pJC96	NS	Liver/807	5.	pJSC33	EF051518	367
6.	pJSC29	EF050083	Spleen/425	6.	pJSC34	EF051519	367
7.	pJC97	NS	Testis/425	7.	pJSC35	NS	277
8.	pJC98	NS	Ovary/425	8.	PJSC36	NS	282
9.	pJC99	NS	Kidney/425	9.	pJSC37	NS	150
10.	pJC100	NS	Liver/425	10.	pJSC38	NS	125
11.	pJC101	NS	Testis/414				
12.	pJC102	NS	Testis/417				
13.	pJC89	EF592585	Testis/376				
14.	pJC103	NS	Ovary/367				
15.	pJC104	NS	Liver/367				
16.	pJC105	NS	Kidney/367				
17.	pJC106	NS	Testis/367				
18.	pJC87	EF592582	Testis/277				
19.	pJC88	EF592583	Testis/282				
20.	pJC107	NS	Ovary/282				
21.	pJC108	NS	Spleen/282				
22.	pJC109	NS	Liver/282				
23.	pJC90	EF592585	Testis/150				
24.	pJC91	EF592586	Testis/125				

The transcripts detected in somatic tissues are described in **(A)** whereas spermatozoal transcripts in **(B)**. Note that these transcripts did not show any homology with genes present in databank.

ones uncovered from testis (Table 12). However, other somatic tissues and ovary shared only 4 out of 10 transcripts with the spermatozoa. Further, homology search and other studies showed that more than 90% of the somatic and spermatozoal transcripts are novel and portraying no/ non-significant homology with any reported genes in the database. Remaining 10% ones were similar to few Bovid specific genes or uncharacterized BAC clones. Interestingly, none of these GATA-tagged transcripts showed homology along their entire length. The details of the somatic and spermatozoal transcripts tagged with GATA repeat including their accession numbers, origin and size are given in the table 12.

4.1.2.4 *Comparative analysis of the transcript diversity*

The observed tissue-specific profiles for the 33.15- (Figure 14A-B) and GACA/GATA- (Figure 15A-B) tagged transcripts so uncovered were further substantiated using RNA slot-blot hybridizations and RT-PCR analyses (Figure 14-15). We next investigated the comparative richness of the buffalo transcriptome for GACA, GATA or 33.15 repeats, and found the association of relatively more number of transcripts with GACA than with GATA or 33.15 repeat. Briefly, 34 different transcripts were identified by GACA encompassing 8 from somatic tissues/gonads, 20 from spermatozoa and 6 shared by all sources studied (Table 11). Contrast to this, only 10 types of transcripts were identified by GATA with 4 common to all the studied sources but 6 exclusive to testis/spermatozoa (Table 12). The consensus of 33.15 repeat uncovered 19 types of mRNA transcripts encompassing 7 each from somatic tissues and gonads while 12 from the spermatozoa.

Most interestingly, none of the 33.15 tagged transcripts was found to be common amongst different tissues studied and the spermatozoa (Table 10). As for the homologous genes, none of them were shared between any of these repeat tagged transcripts. Moreover, all the homologues were found to be either involved in signal transduction/other regulatory pathways or linked to several diseases. Based on these findings, we conclude that GACA identified more number of known and novel genes whereas GATA uncovered fewer but only novel ones.

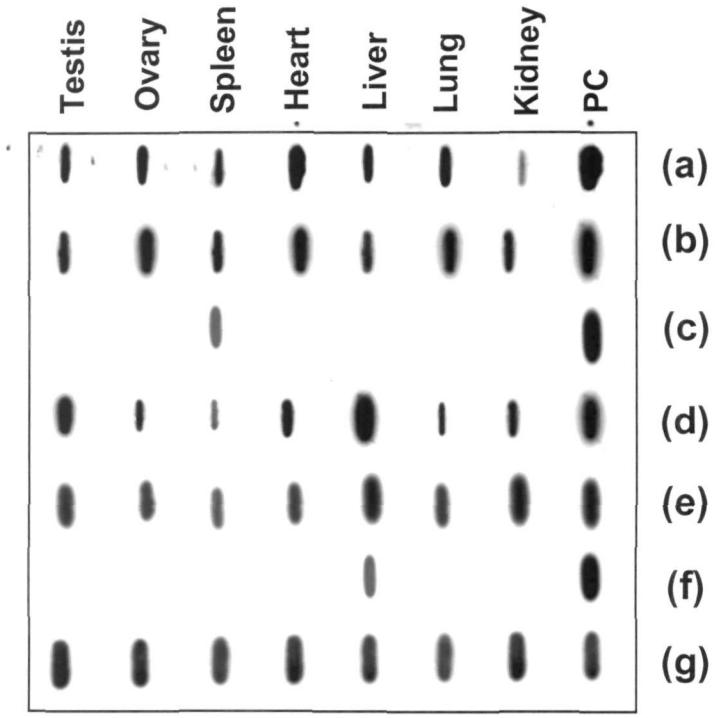


Figure 14A. Slot blot hybridization of total RNA from different tissues with cloned probes having insert fragments of: 324, **(a)**; 487, **(b)**; 576, **(c)**; 602, **(d)**; 846 **(e)**; 1263 **(f)** and β-actin **(g)**. Note the exclusive signals detected in liver and spleen by 1263 and 576 bp cloned fragments, respectively. PC denotes 5 ng respective recombinant plasmids used as positive control.

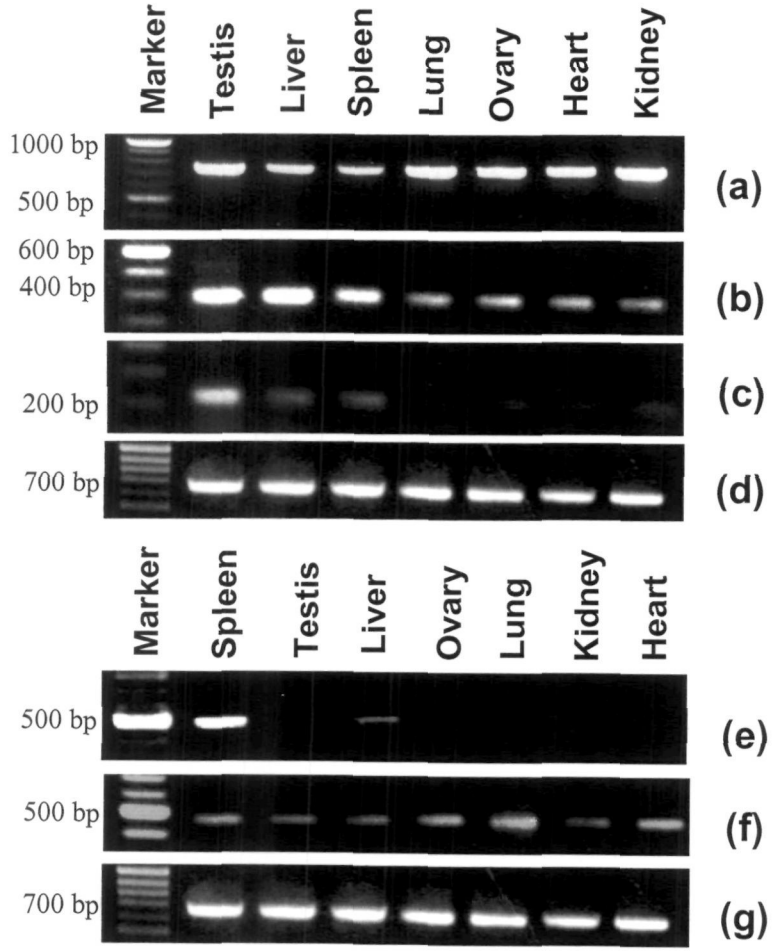


Figure 14B. RT-PCR amplification of cDNA templates from different tissues showing amplicons corresponding to the insert fragments: 846 **(a)**; 487 (**b**); 324 **(c)**; β-actin **(d)**; 576 **(e)**; 602 **(f)** and β-actin **(g)**. Note the faint signals of 324 related fragment in lung, liver, heart and kidney (panel **c**) and the absence of 576 base pair related signal in testis, ovary, lung, kidney and heart (panel e).

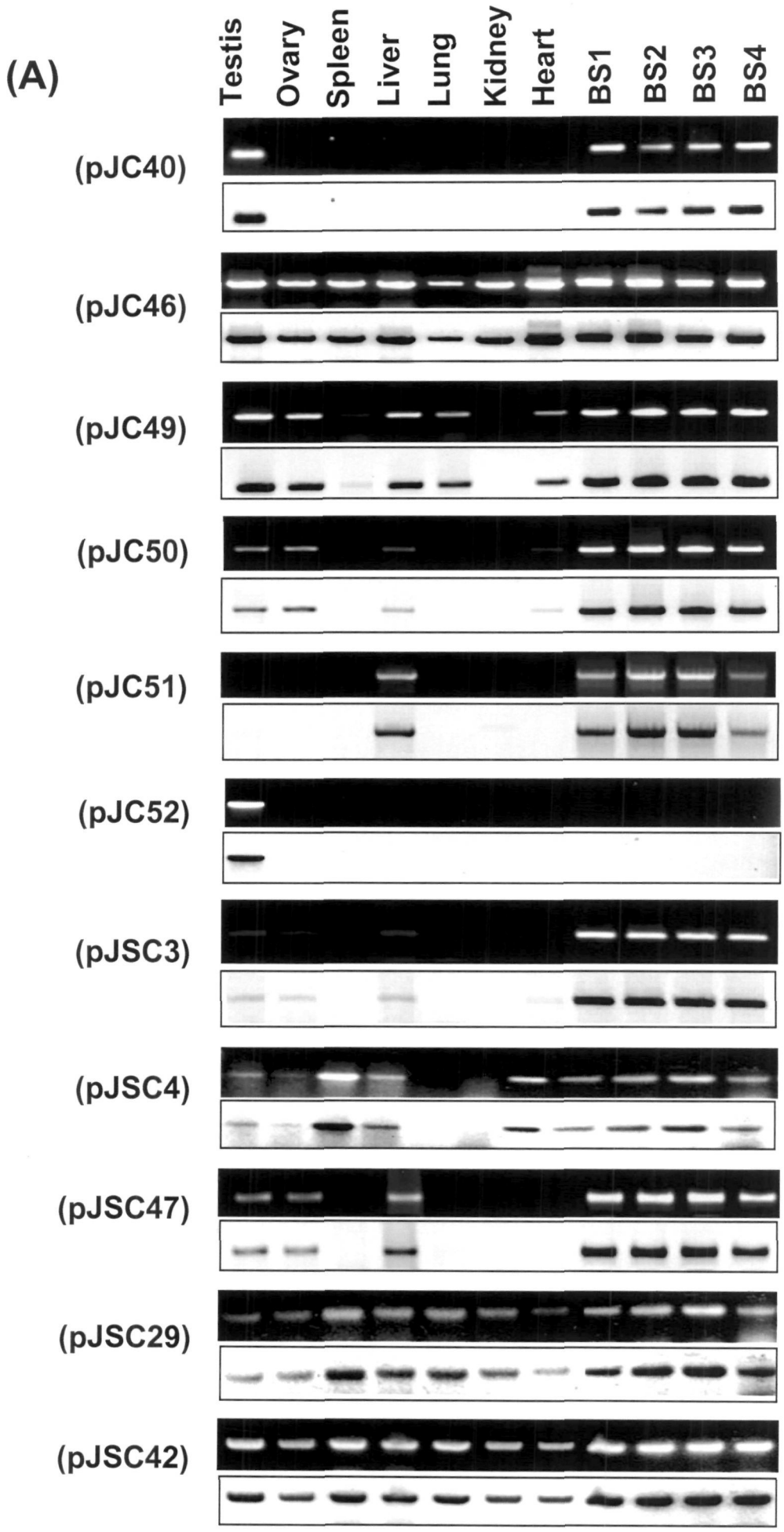


Figure 15

Contd/-

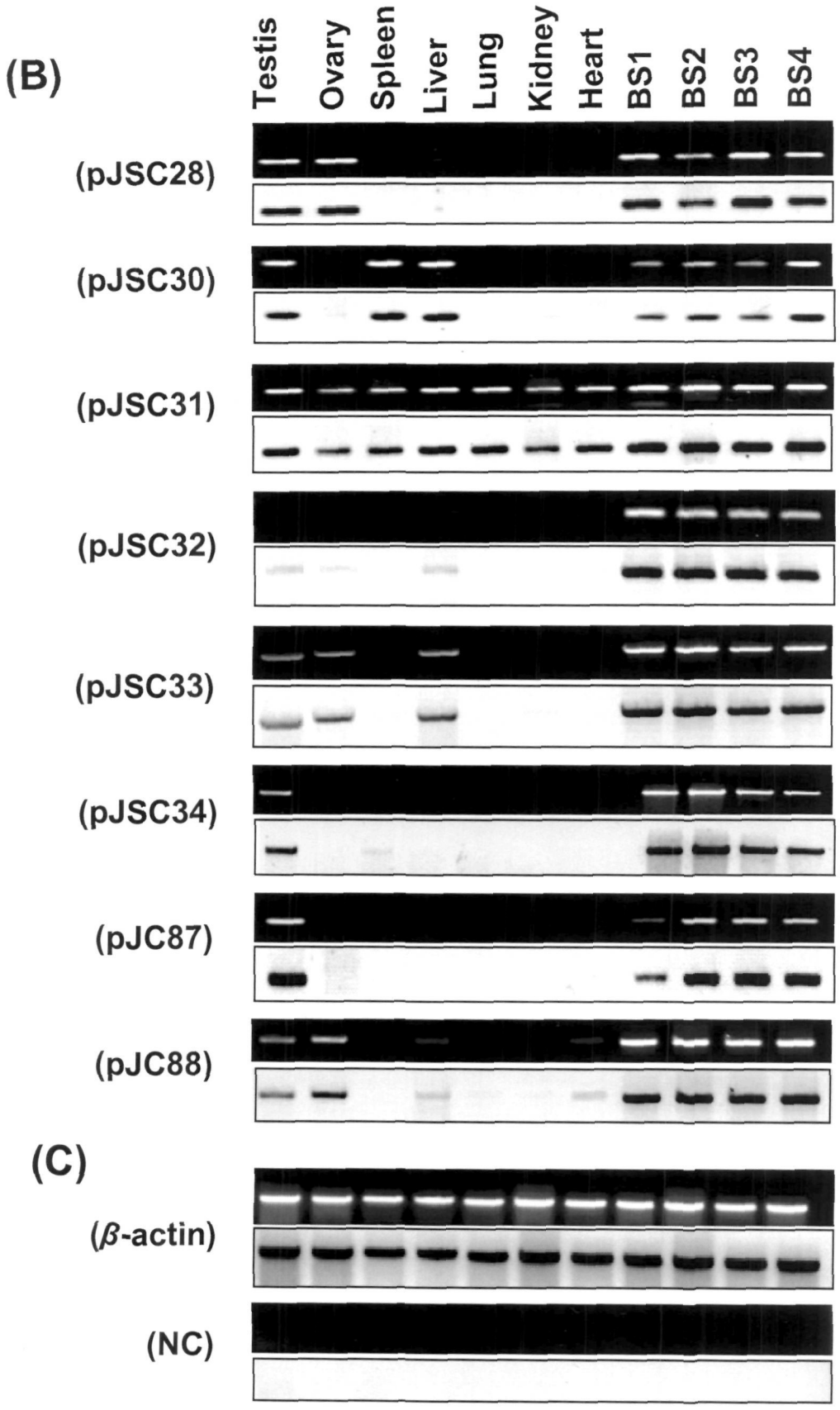


Figure 15. RT-PCR analyses for representative GACA- **(A)** and GATA- **(B)** tagged transcripts using internal primers and cDNA from different somatic tissues, gonads and spermatozoa as templates. The transcript IDs are given on the left and names of the tissues on the top. Quality and quantity of the cDNA samples was normalized **(C)** and genomic contamination in the RNA checked by PCR with β-actin derived primers. Tissue specificities of the transcripts were ascertained on the basis of presence or absence of amplicons using the respective cDNA templates which were further confirmed by Real Time PCR and Southern blotting.

Moreover, 33.15 repeat also detected less transcripts than the GACA repeat but more than the GATA one.

4.1.3 Sequence polymorphisms detected in the MASA uncovered transcripts

Following homology search, we analyzed all the transcripts for sequence polymorphisms at inter-tissue or tissue-spermatozoal levels. The possibility of inter-clonal sequence variations for the individual fragment was ruled out as described in the section 4.1.2. In brief, a total of 148 amplicons were uncovered with 33.15 repeat comprising 100 from somatic and gonadal tissues and 48 from spermatozoa from four different animals which were cloned and sequenced (Table 10). Also, 332 amplicons encompassing 57 from somatic tissues/gonads and 26 from spermatozoa, each from 4 animals were uncovered with GACA (Table 11), and 136 amplicons encompassing 96 from different tissues and 40 from spermatozoa tagged with GATA were cloned and sequenced (Table 12).

4.1.3.1 *Within the 33.15 tagged transcripts*

As mentioned above, no transcript was found to be shared between the tissues and the spermatozoa. Among the transcripts exclusively detected in the somatic tissues/gonads, the 846/847 bp fragment showed tissue specific point nucleotide changes when analyzed using ClustalW alignment (www.ebi.ac.uk/clustalw). A total of 6 changes specific to ovary and testis and 6 random ones in spleen, liver, lung, kidney and heart were detected (Figure 16A). *In silico* translation of these sequences showed amino acids changes leading to massive C-terminus alterations and deletions in gonads of both the sexes compared to that in somatic tissues (Figure 16B). Startlingly, this change was not confined to one animal, instead was uncovered in others too. Sequences of 602 bp fragment, which was identical in testis, ovary, kidney, liver and lung showed 2 point mutations in heart. In spleen a 576 bp fragment representing altogether different sequence was found instead of the 602 bp one (Figure 17). Interestingly, the 487 bp fragment, which was found to homologous to the palindromic sequence of the Y chromosome, had identical sequences in

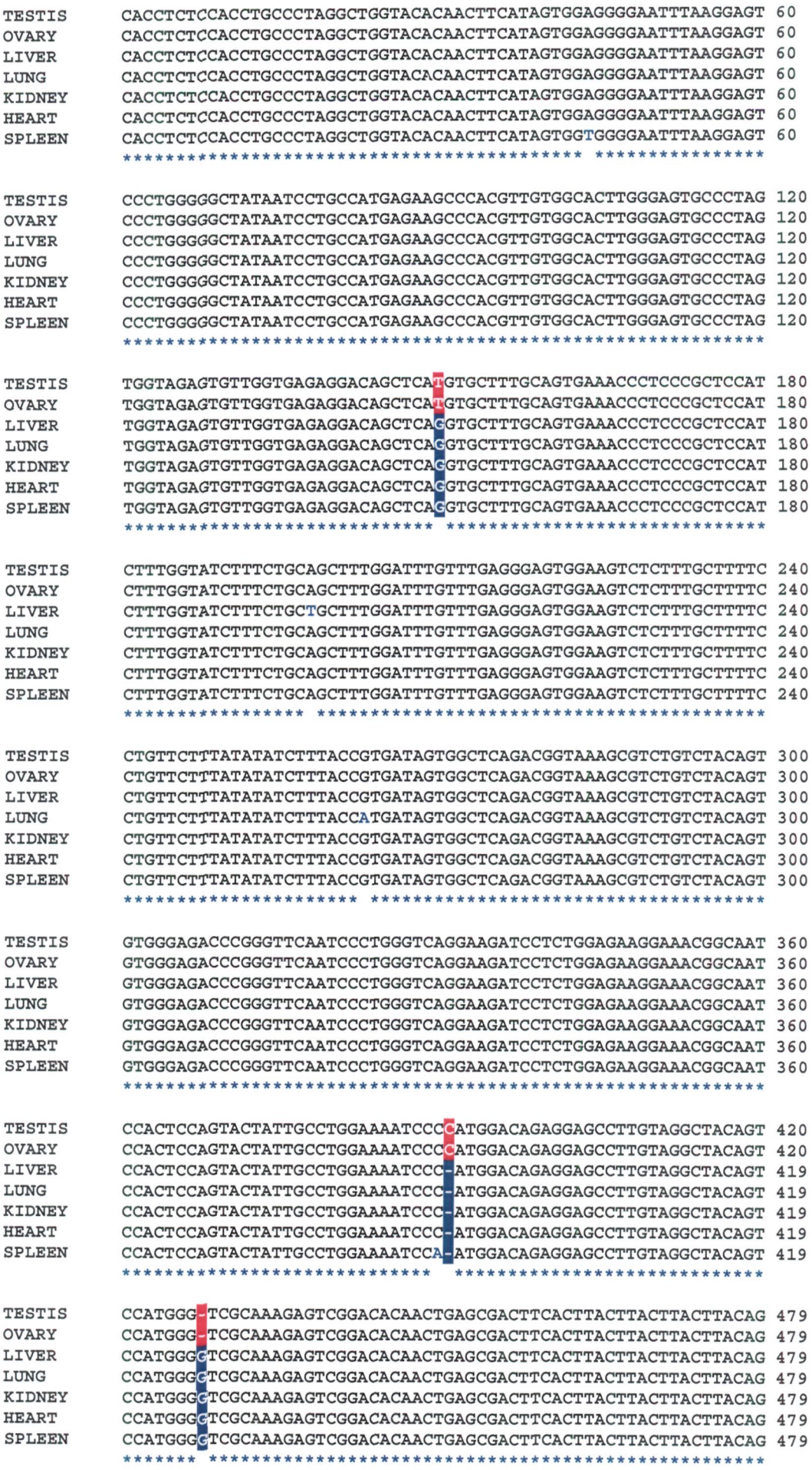


Figure 16A

Contd/-

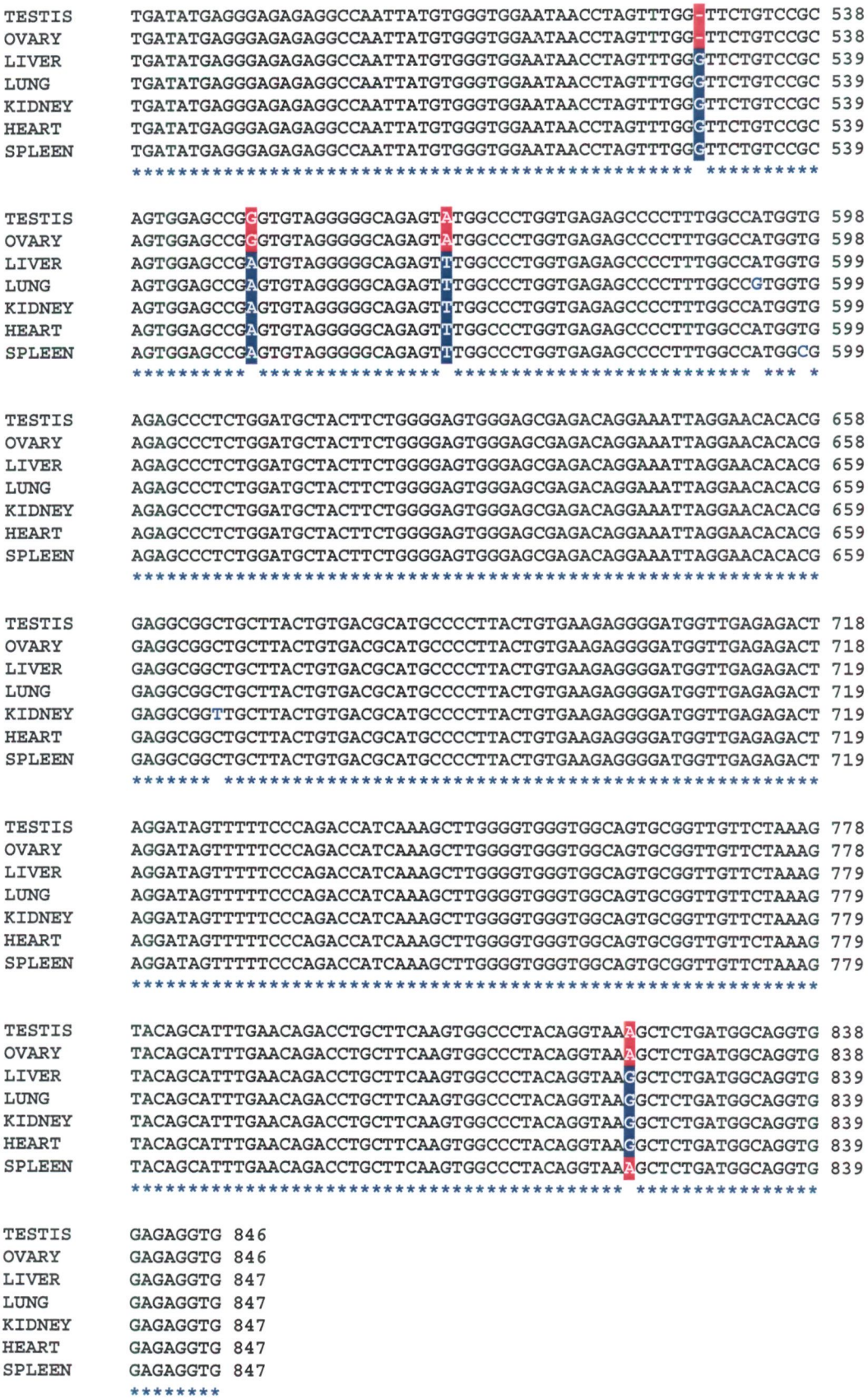


Figure 16A. Multiple sequence alignment of 846/847 bp nucleotide sequences (similar to Adenylate kinase) from seven different tissues showing identical changes at six different positions in testis and ovary (gonad specific) compared with that in somatic tissues.

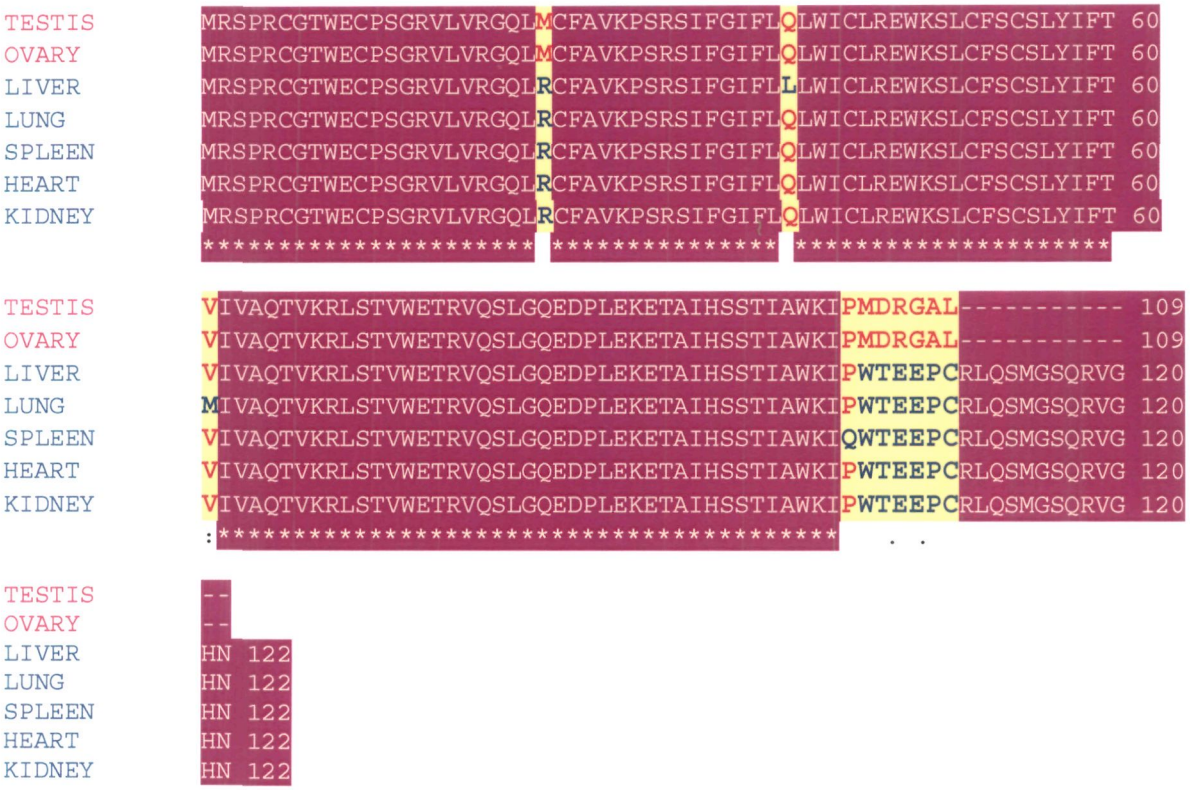


Figure 16B. Multiple sequence alignment of amino acid sequences deduced from 846/847 transcript, originating from seven different tissues. Please note the conspicuous C-terminal alterations and deletions of the residues in the gonads compared to that in somatic tissues.

```
LUNG          CACCTCTCCACCTGCCTCCAAGTCAAGGTGAGTTAGGAGTTAGATTTCCTCCTTTATGTA 60
OVARY         CACCTCTCCACCTGCCTCCAAGTCAAGGTGAGTTAGGAGTTAGATTTCCTCCTTTATGTA 60
LIVER         CACCTCTCCACCTGCCTCCAAGTCAAGGTGAGTTAGGAGTTAGATTTCCTCCTTTATGTA 60
HEART         CACCTCTCCACCTGCCTCCAAGTCAAGGTGAGTTAGGAGTTAGATTTCCTCCTTTATGTA 60
TESTIS        CACCTCTCCACCTGCCTCCAAGTCAAGGTGAGTTAGGAGTTAGATTTCCTCCTTTATGTA 60
KIDNEY        CACCTCTCCACCTGCCTCCAAGTCAAGGTGAGTTAGGAGTTAGATTTCCTCCTTTATGTA 60
              ************************************************************
SPLEEN        CACCTCTCCACCTGCCCCCCAG--AGCTTCCCCCAGGTGCCTCTGGCTTGTCTCTGGCGA 58
              *************** ** **  *   *     *** *           ** *    *

LUNG          TTCCTTTGGACATATGATCATTATCATGGTTGGGTTTTCTTGTCACAGTCTCTTCCAGCA 120
OVARY         TTCCTTTGGACATATGATCATTATCATGGTTGGGTTTTCTTGTCACAGTCTCTTCCAGCA 120
LIVER         TTCCTTTGGACATATGATCATTATCATGGTTGGGTTTTCTTGTCACAGTCTCTTCCAGCA 120
HEART         TTCCTTTGGACATATGATCATTATCATGGTTGGGTTTTCTTGTCACAGTCTCTTCCAGCA 120
TESTIS        TTCCTTTGGACATATGATCATTATCATGGTTGGGTTTTCTTGTCACAGTCTCTTCCAGCA 120
KIDNEY        TTCCTTTGGACATATGATCATTATCATGGTTGGGTTTTCTTGTCACAGTCTCTTCCAGCA 120
              ************************************************************
SPLEEN        CACCTCTGGTTCCTC--CCGTTCTCCTCATAGGCTCCTCAGAGCCTGGGGCCTTCC---- 112
                *** ***         * ** ** *  * ** *  **    *   *   *****

LUNG          GACTTAAAAGATCAAGAGAAGATGGACTATCTTATTAAAACTCTGGATTCTTAATGTGCA 180
OVARY         GACTTAAAAGATCAAGAGAAGATGGACTATCTTATTAAAACTCTGGATTCTTAATGTGCA 180
LIVER         GACTTAAAAGATCAAGAGAAGATGGACTATCTTATTAAAACTCTGGATTCTTAATGTGCA 180
HEART         GACTTAAAAGATCAAGAGAAGATGGACTATCTTATTAAAACTCTGGATTCTTAATGTGCA 180
TESTIS        GACTTAAAAGATCAAGAGAAGATGGACTATCTTATTAAAACTCTGGATTCTTAATGTGCA 180
KIDNEY        GACTTAAAAGATCAAGAGAAGATGGACTATCTTATTAAAACTCTGGATTCTTAATGTGCA 180
              ************************************************************
SPLEEN        ---TCAAGAGGCTTCCTGAGCCCAGTCTCCCCCGCCTATCCCCAAGGTT----GTGAACA 165
                 * ** **       **     * **  *      *  * *  * **     **  **

LUNG          CAGTGCCTGCACATGACACAGTGGTCAATAAATGCTTGTTGAATGTGTGATTATATAATA 240
OVARY         CAGTGCCTGCACATGACACAGTGGTCAATAAATGCTTGTTGAATGTGTGATTATATAATA 240
LIVER         CAGTGCCTGCACATGACACAGTGGTCAATAAATGCTTGTTGAATGTGTGATTATATAATA 240
HEART         CAGTGCCTGCACATGACACAGTGGTCAATAAATGCTTGTTGAATGTGTGATTATATAATA 240
TESTIS        CAGTGCCTGCACATGACACAGTGGTCAATAAATGCTTGTTGAATGTGTGATTATATAATA 240
KIDNEY        CAGTGCCTGCACATGACACAGTGGTCAATAAATGCTTGTTGAATGTGTGATTATATAATA 240
              ************************************************************
SPLEEN        CCCTGAAAGCACA--GCATTTTGGT--ATCCCTGCCTTGCACCCAGTCCCTCACTCAGGG 221
              *  **   *****   **   ****  **   *** *            * *   *

LUNG          TTTTGAGCAATAAAATCTTAAATTAGAAGATAATTTGCCCCTTGATTCATACAGTATACT 300
OVARY         TTTTGAGCAATAAAATCTTAAATTAGAAGATAATTTGCCCCTTGATTCATACAGTATACT 300
LIVER         TTTTGAGCAATAAAATCTTAAATTAGAAGATAATTTGCCCCTTGATTCATACAGTATACT 300
HEART         TTTTGAGCAATAAAATCTTAAATTAGAAGATAATTTGCCCCTTGATTCATACAGTATACT 300
TESTIS        TTTTGAGCAATAAAATCTTAAATTAGAAGATAATTTGCCCCTTGATTCATACAGTATACT 300
KIDNEY        TTTTGAGCAATAAAATCTTAAATTAGAAGATAATTTGCCCCTTGATTCATACAGTATACT 300
              ************************************************************
SPLEEN        CAGTAAACAGTAGGTGCTCAATTTGGATG-TGAACTATTCACAGAAGAAGCGAGTGAAAG 280
                 * * ** **    ** ** ** ** * * *  *   *   **   *   ***  *

LUNG          CGACAAATTAAGAAGCCTTTCAACAGACTGAAGAAAGACAGCAAGATTGTAAAGTTACAG 360
OVARY         CGACAAATTAAGAAGCCTTTCAACAGACTGAAGAAAGACAGCAAGATTGTAAAGTTACAG 360
LIVER         CGACAAATTAAGAAGCCTTTCAACAGACTGAAGAAAGACAGCAAGATTGTAAAGTTACAG 360
HEART         CGACAAATTAAGAAGCCTTTCAACAGACTGAAGAAAGACAGCAAGATTGTAAAGTTACAG 360
TESTIS        CGACAAATTAAGAAGCCTTTCAACAGACTGAAGAAAGACAGCAAGATTGTAAAGTTACAG 360
KIDNEY        CGACAAATTAAGAAGCCTTTCAACAGACTGAAGAAAGACAGCAAGATTGTAAAGTTACAG 360
              ************************************************************
SPLEEN        TGAAAGTGAATGTCACTCAGTCATGTCCAACTCTTTGCAACTCCATGGACTATACAGTCC 340
               ** *    * *   *    *     *  ***      *        ** * ** ***
```

Figure 17

Contd/

```
LUNG        CAGATTCCTTCTAGATCAAACAGACTGCTTTCTCTCTAAGCCTGGGTGGTGGTGGTAGTG 420
OVARY       CAGATTCCTTCTAGATCAAACAGACTGCTTTCTCTCTAAGCCTGGGTGGTGGTGGTAGTG 420
LIVER       CAGATTCCTTCTAGATCAAACAGACTGCTTTCTCTCTAAGCCTGGGTGGTGGTGGTAGTG 420
HEART       CAGATTCCTTCTAGATCAAACAGACTGCTTTCTCTCTAAGCCTGGGTGGTGGTGGTAGTG 420
TESTIS      CAGATTCCTTCTAGATCAAACAGACTGCTTTCTCTCTAAGCCTGGGTGGTGGTGGTAGTG 420
KIDNEY      CAGATTCCTTCTAGATCAAACAGACTGCTTTCTCTCTAAGCCTGGGTGGTGGTGGTAGTG 420
            ************************************************************
SPLEEN      ATGGAATTTTCCAGGCCAAAATAGTGGAGTGGGTAGCCTTTCCCTTCTCCATGGG--ATC 398
              * *  ** *    * * *  *    *   *   ** **  *  *   * ****   *

LUNG        TTGCTGTTCAGTTGCAAGTCATGTCTGACTCTTGAGACCCCATGGACTGTAGCCCACCAG 480
OVARY       TTGCTGTTCAGTTGCAAGTCATGTCTGACTCTTGAGACCCCATGGACTGTAGCCCACCAG 480
LIVER       TTGCTGTTCAGTTGCAAGTCATGTCTGACTCTTGAGACCCCATGGACTGTAGCCCACCAG 480
HEART       TTGCTGTTCAGTTGCAAGTCATGTCTGACTCTTGAGACCCCATGGACTGTAGCCCACCAG 480
TESTIS      TTGCTGTTCAGTTGCAAGTCATGTCTGACTCTTGAGACCCCATGGACTGTAGCCCACCAG 480
KIDNEY      TTGCTGTTCAGTTGCAAGTCATGTCTGACTCTTGAGACCCCATGGACTGTAGCCCACCAG 480
            ************************************************************
SPLEEN      TTCCCAAT-TCAGGGATCAAACCCAGGTCTCCCGCACCGTGGGCGGATTCTTTACCAGCT 457
             *   * * *    * ***    ** ***        *    * *    * *   * *

LUNG        GCTTCTCCAGCGATTGGATTTCTCAGTAAGAATACTGGAGTGGACTGCCATTTCTTTCTC 540
OVARY       GCTTCTCCAGCGATTGGATTTCTCAGTAAGAATACTGGAGTGGACTGCCATTTCTTTCTC 540
LIVER       GCTTCTCCAGCGATTGGATTTCTCAGTAAGAATACTGGAGTGGACTGCCATTTCTTTCTC 540
HEART       GCTTCTCCAGCGATTGGATTTCTCAGTAAGAATACTGGAGTGGACTGCCATTTCTTTCTC 540
TESTIS      GCTTCTCCAGCGATTGGATTTCTCAGTAAGAATACTGGAGTGGACTGCCATTTCTTTCTC 540
KIDNEY      GCTTCTCCAGCGATTGGATTTCTCAGTAAGAATACTGGAGTGGACTGCCATCTCCTTCTC 540
            ********************************************** ** *****
SPLEEN      GAGCCACAAG-TGAAGAAGTGA--GGGATGCTCAGTAAAAGCCTATGGAAGAGTAGGTGG 514
            *  *  ** * *     ** **  **  * ******** *      *   *    *

LUNG        TGGGGGATCCTCCCAACCCAAGAATTGAATTTGCATTTCTTGCATTGGCAGGTGGAGAGG 600
OVARY       TGGGGGATCCTCCCAACCCAAGAATTGAATTTGCATTTCTTGCATTGGCAGGTGGAGAGG 600
LIVER       TGGGGGATCCTCCCAACCCAAGAATTGAATTTGCATTTCTTGCATTGGCAGGTGGAGAGG 600
HEART       TGGGGGATCCTCCCAACCCAAGAATTGAATTTGCATTTCTTGCATTGGCAGGTGGAGAGG 600
TESTIS      TGGGGGATCCTCCCAACCCAAGAATTGAATTTGCATTTCTTGCATTGGCAGGTGGAGAGG 600
KIDNEY      CGGGGGATCCTCCCAACCCAAGAATTGAATTTGCATTTCTTGCATTGGCAGGTGGAGAGG 600
             ************************************************************
SPLEEN      TTATAGTTCAAGGG-CCCTTACTCTG---TTCTGCTCTC-TTCTCAGGCAGGTGGAGAGG 565
            ** *   **  ***      **  ** *     **    *  * **  ***********

LUNG        TG 602
OVARY       TG 602
LIVER       TG 602
HEART       TG 602
TESTIS      TG 602
KIDNEY      TG 602
SPLEEN      TG 576
            **
```

Figure 17. Multiple alignment of 576 bp fragment from spleen with 602 bp ones from the other tissues. Note altogether different sequence in 576 bp fragment.

all the tissues including that in ovary (Figure 18). Similarly, the 324 bp fragment, though detected only in lung, testis and spleen also showed identical sequences (not shown).

4.1.3.2 *In the GACA tagged transcripts*

Our study demonstrated several single nucleotide changes and INDEL polymorphisms in most of the GACA-tagged transcripts. As mentioned above, only 9 transcripts were common amongst tissues and spermatozoa with the remaining ones restricted to a single tissue or sperm. For instance, among the transcripts detected exclusively in the different tissues, the 1.8 kb transcript (GenBank Accession numbers: DQ289479-DQ289486) depicted major alterations including insertions of 36 and 4 bp exclusively in lung and several point nucleotide changes in lung/heart or testis/ovary besides a few randomly distributed ones (Figure 19). The transcripts shared by spermatozoa and tissues also brought out some interesting features. For instance, the 1.3 kb transcript (GenBank Accession numbers DQ534902 and DQ534903) showing homology with NFATC2 gene demonstrated the insertion of 10 bp and several single-nucleotide variations exclusively in spermatozoa (Figure 20).

Next, the point nucleotide changes detected in transcript similar to HBGF-1 gene (GenBank Accession no. DQ534904) were either common to different tissues, or in the spermatozoa (Figure 21). Similarly, random deletions, insertions and transversion/transition at various points of 635 bp transcript similar to WASF2 gene, were detected only in testis compared to that in other tissues and spermatozoa (Figure 22). Interestingly, the 550 bp Ankyrin repeat domain-26 (GenBank Accession number: DQ534906) showed identical nucleotide sequences in both the testis and sperm, but polymorphism at several points in the ovary. This transcript was not detected in any of the somatic tissues (Figure 23). Remaining transcripts such as β-transducin repeat and the novel 450 bp and 209 bp ones (Table 11) showed identical sequences amongst the tissues except for a few point nucleotide changes (not shown).

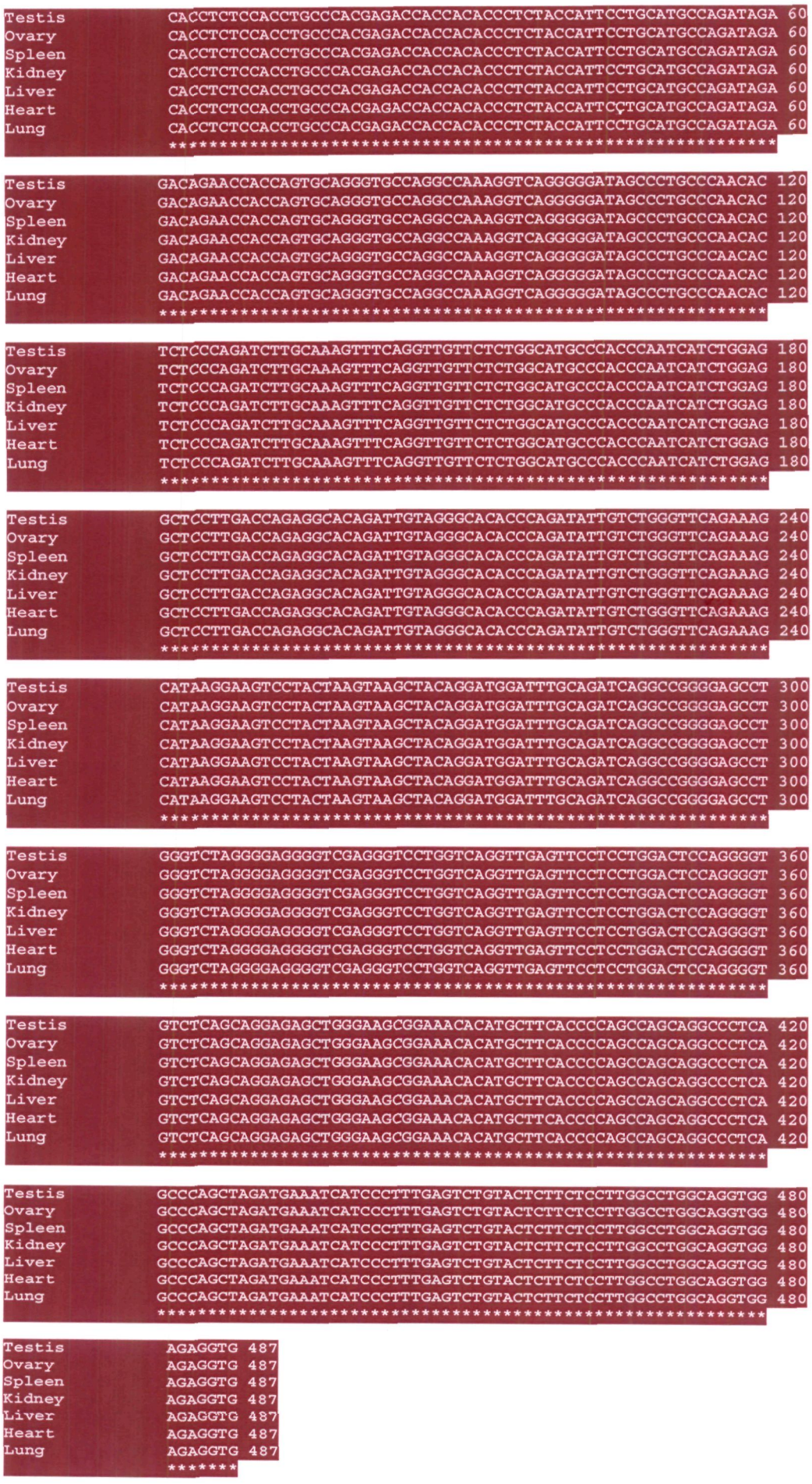


Figure 18. Multiple nucleotide sequence alignment showing conservation of 487 bp fragment across the tissues.

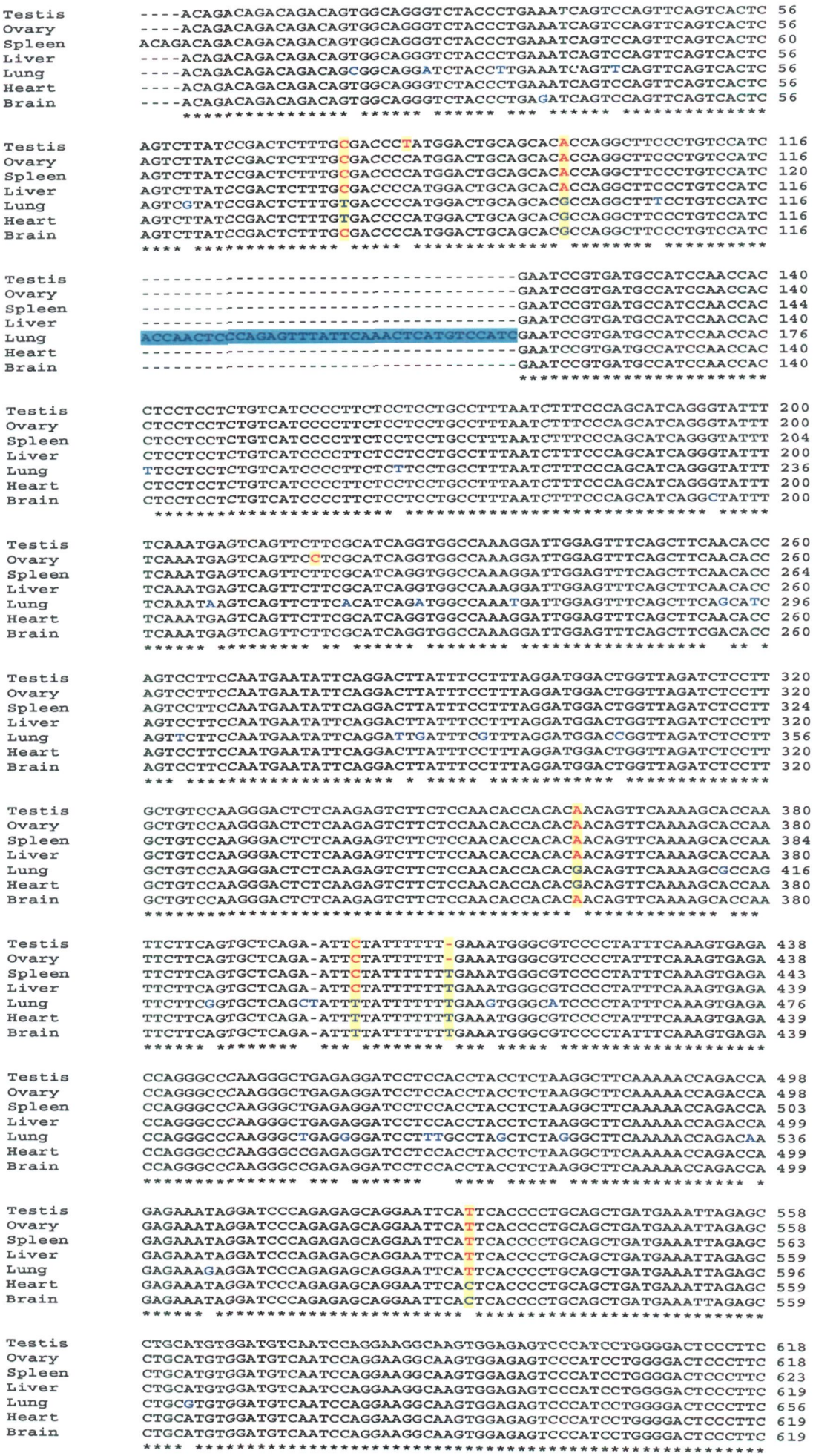


Figure 19

Contd/-

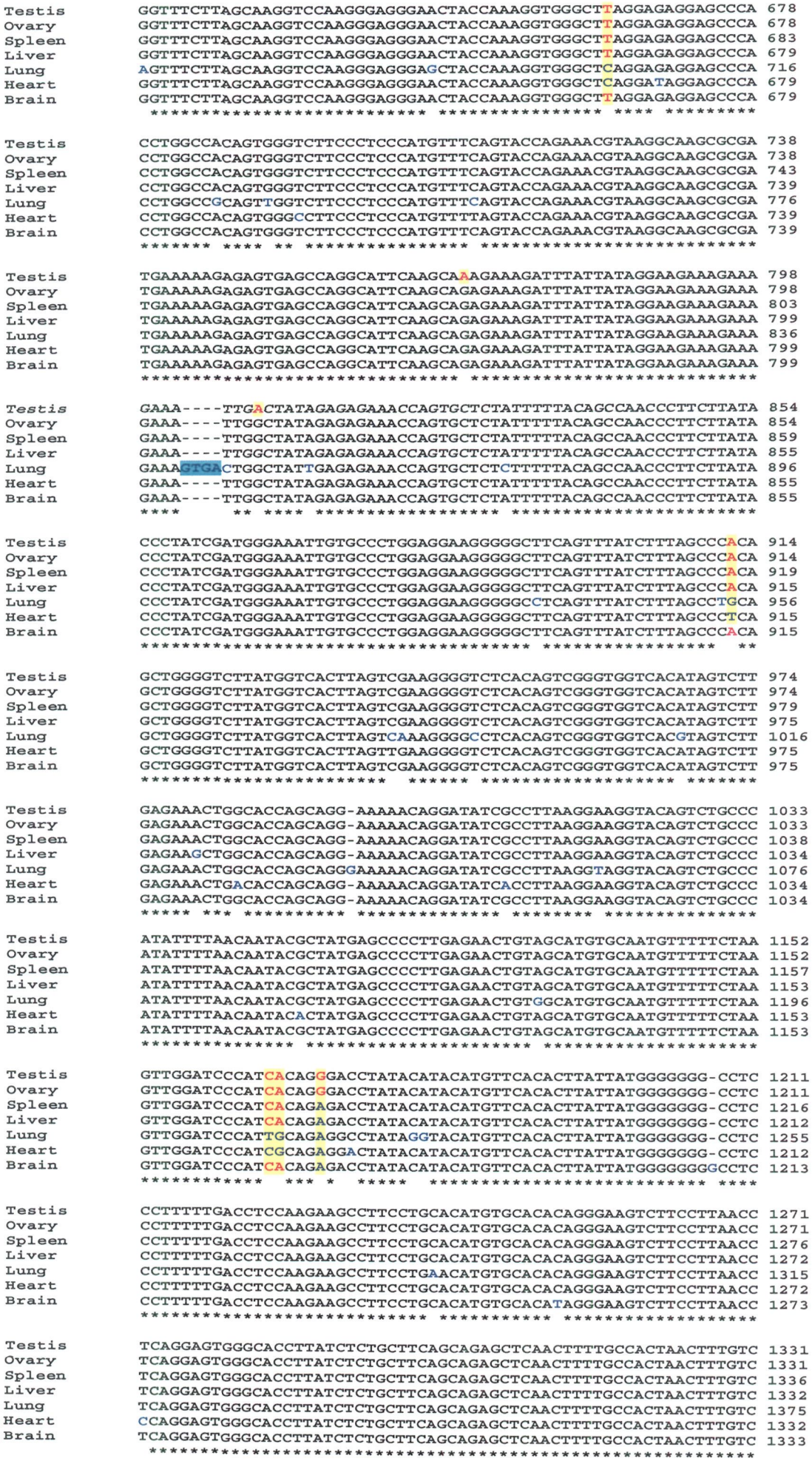


Figure 19

Contd/-

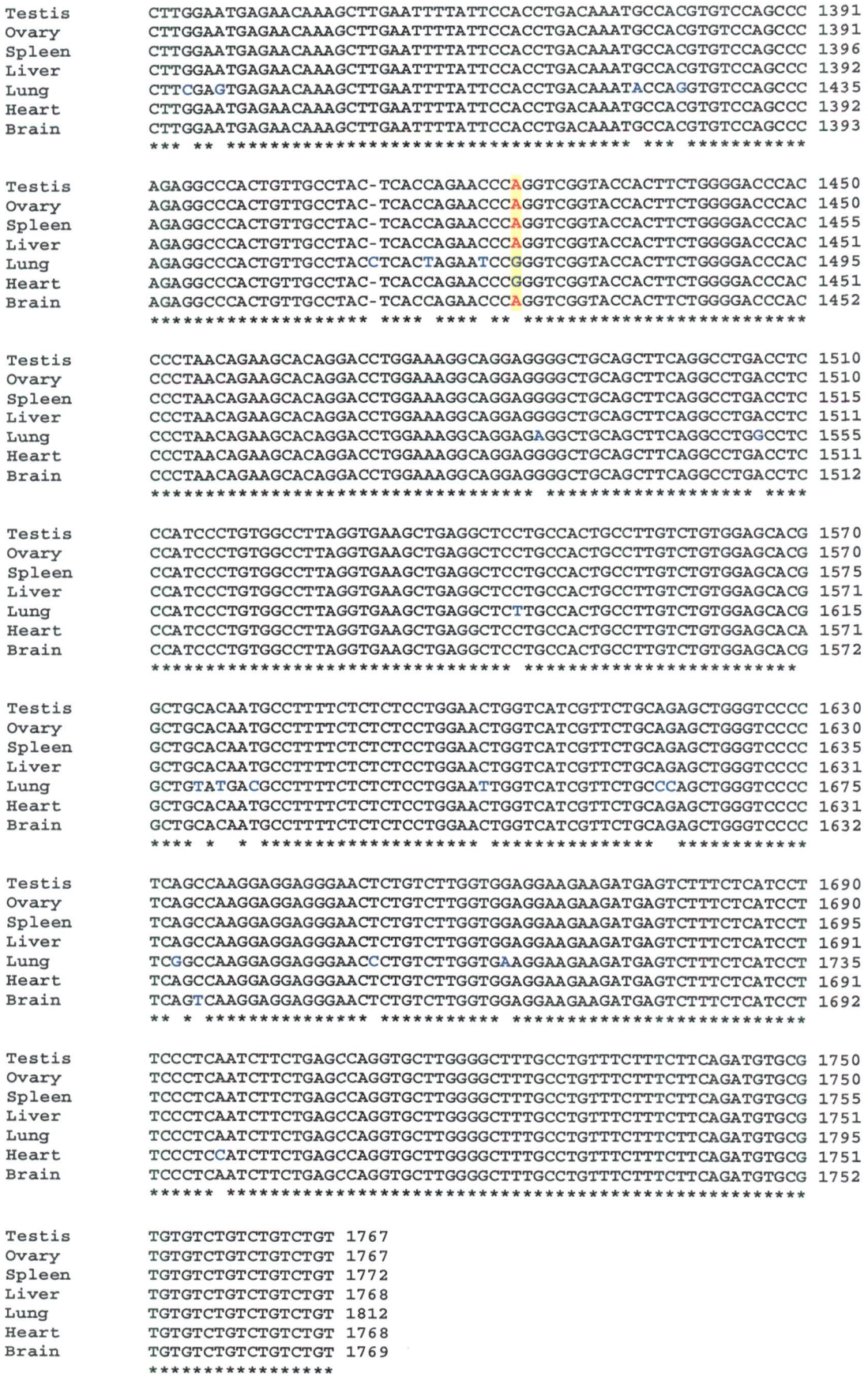


Figure 19. Multiple sequence alignment of GACA-tagged 1.8 kb transcripts from different somatic and gonadal tissues. Note the single nucleotide variations throughout the sequence. The variations shared by gonads and somatic tissues are highlighted in red, the ones common to somatic tissues in blue and gonad specific in pink. Note the exclusive major insertions of 36 and 5 bp in lung, highlighted in blue background, which were reconfirmed by sequencing this fragment from 5 different animals.

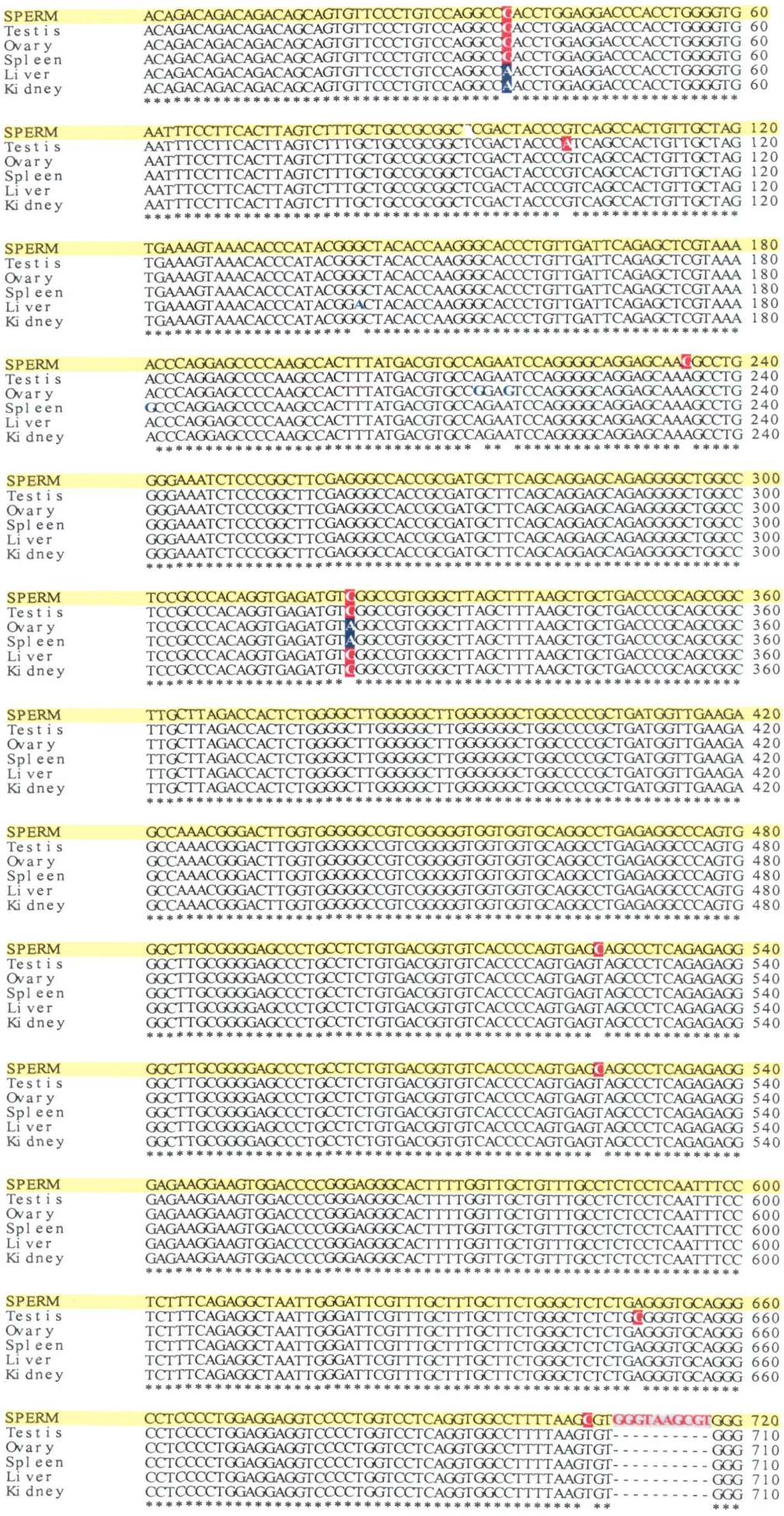


Figure 20

Contd/-

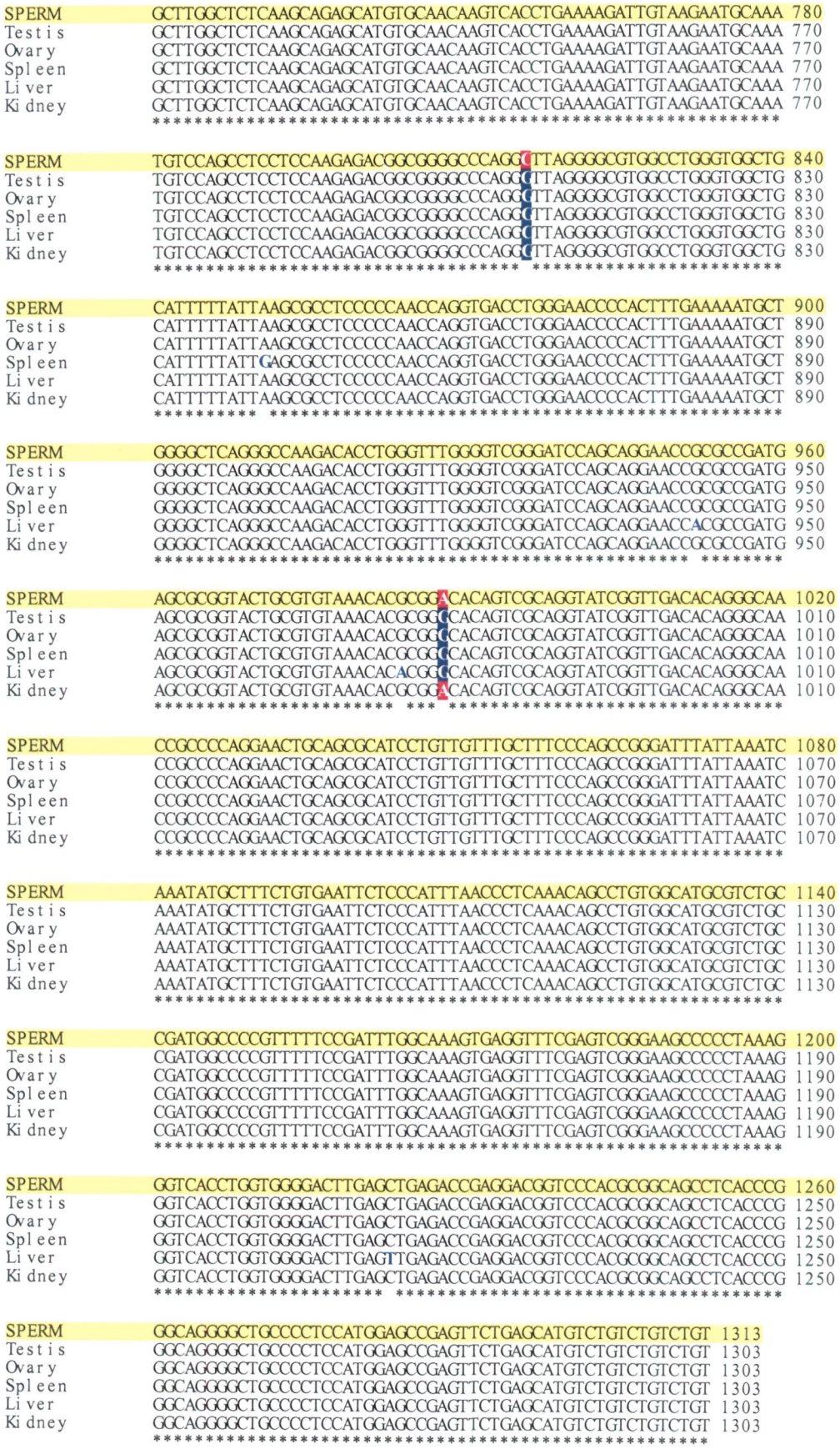


Figure 20. Multiple nucleotide sequence alignment of GACA-tagged 1.3 kb transcripts originating from different tissues and spermatozoa. The sequence from spermatozoa is highlighted in yellow background. The single nucleotide variations spread along the sequence shared by sperm and other tissues are highlighted in pink, and the ones common to tissues in blue. Several variations detected in sperm or testis only are shown in red. Note the exclusive insertion of 10 bp detected in sperm, highlighted in bold red and grey background.

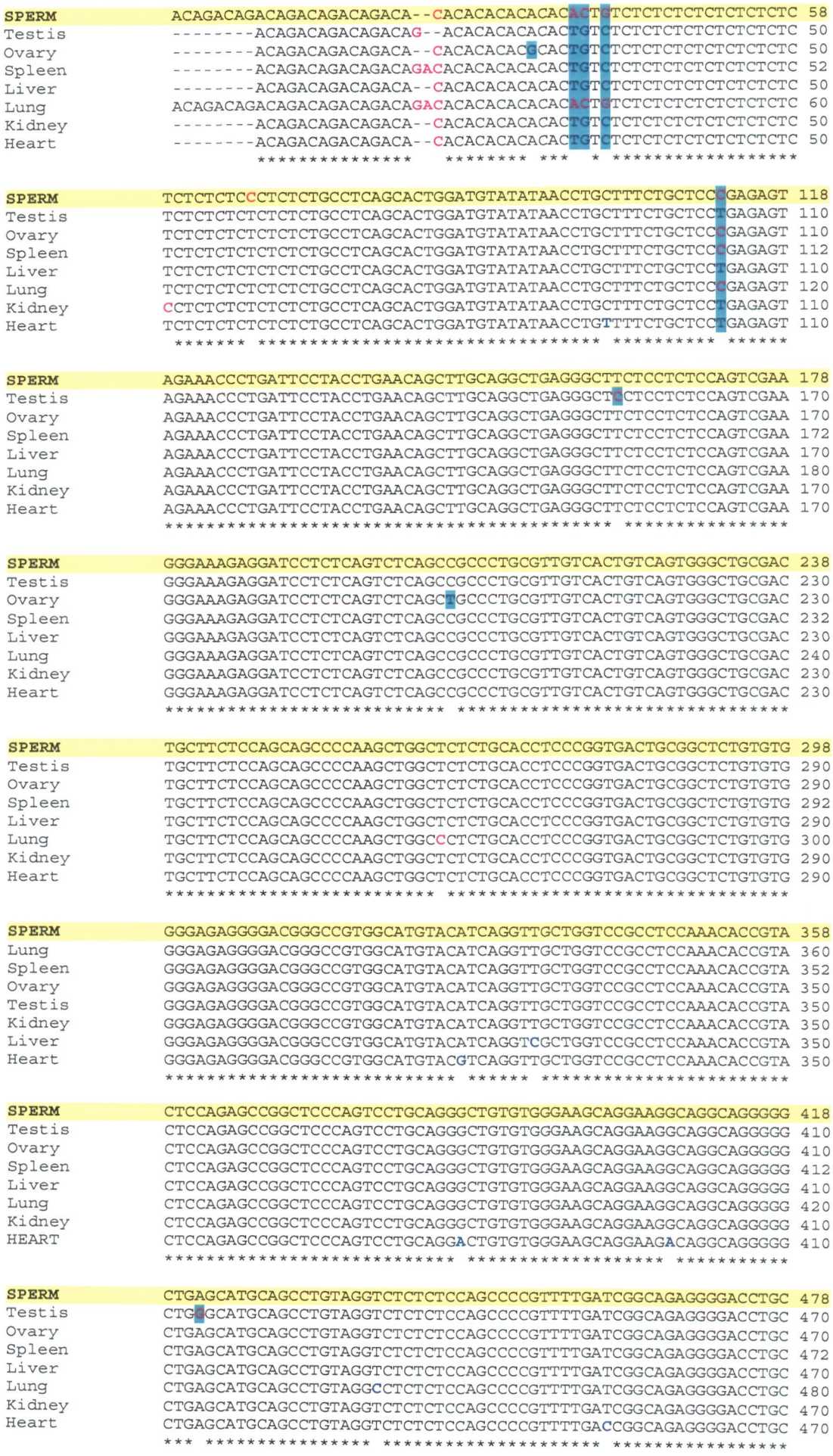


Figure 21

Contd/-

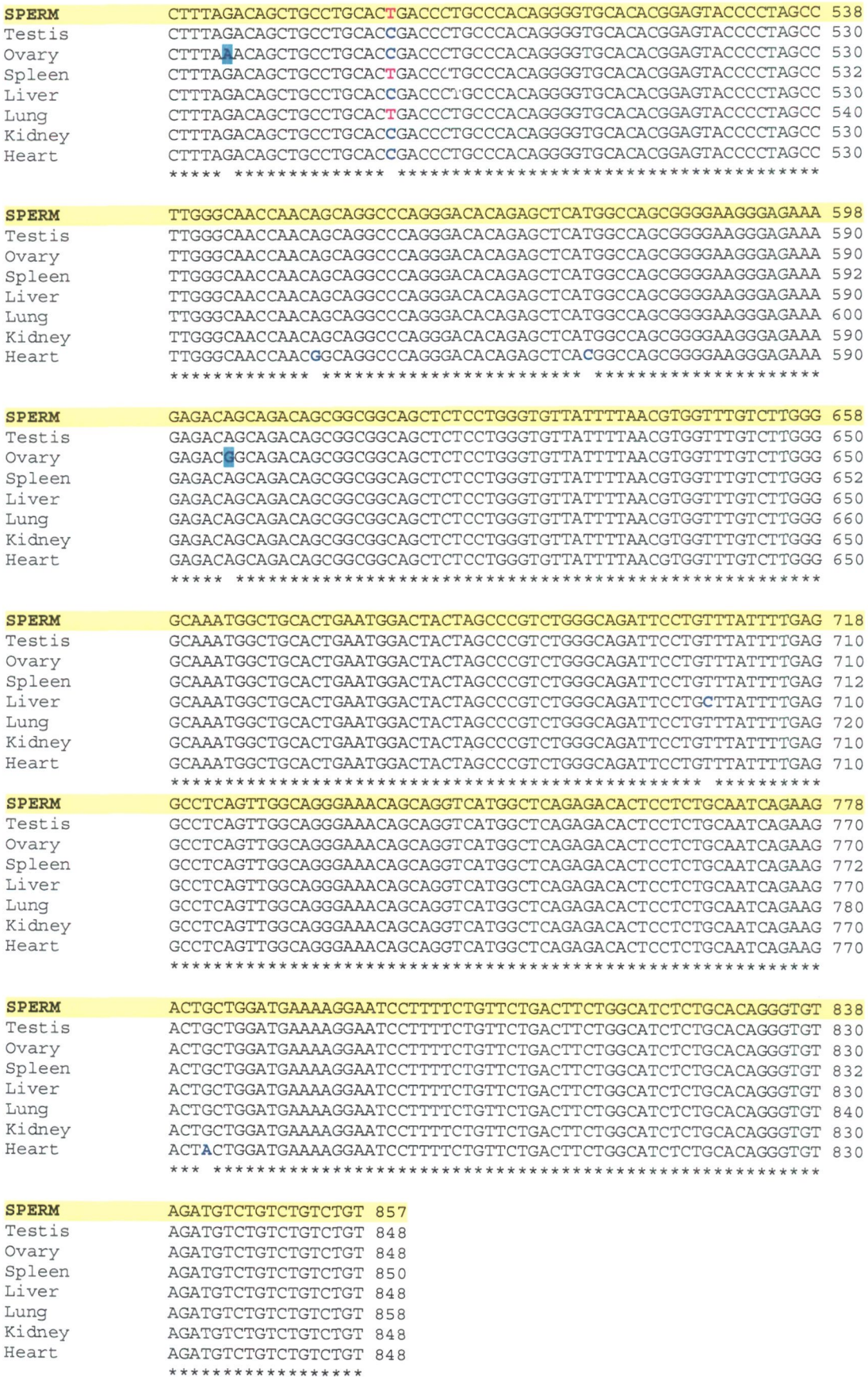


Figure 21. Multiple alignment of GACA-tagged 850 bp transcript originating from different tissues and spermatozoa, homologous to HBGF-1. The sequence from the spermatozoa is highlighted in yellow background. The single nucleotide variations common across the tissues are highlighted in same color (pink or blue). Several variations detected in sperm or testis only are shown in red, and the ones exclusive to ovary in the blue background.

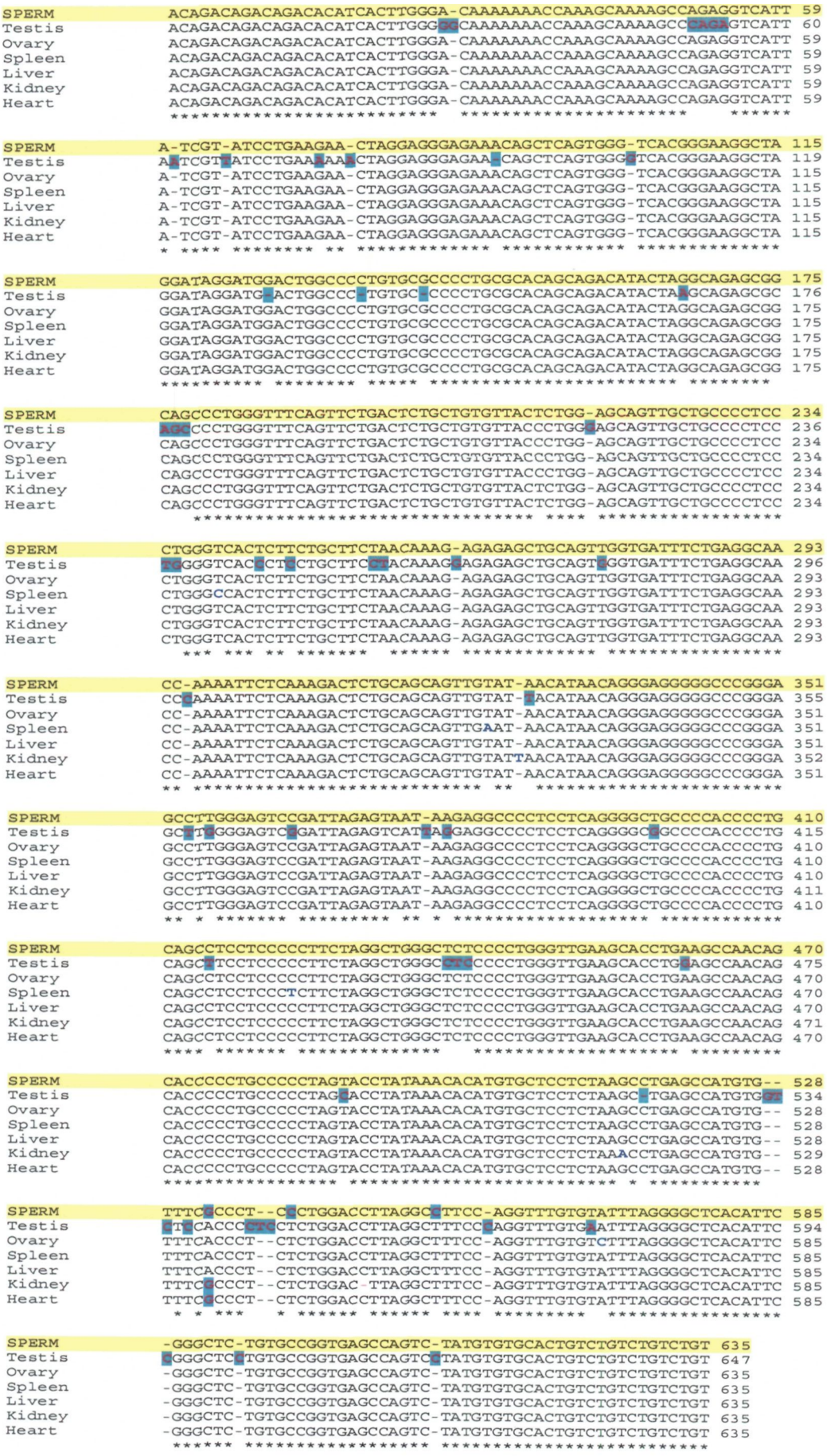


Figure 22. Multiple alignment of GACA-tagged 635 bp transcript originating from spermatozoa and different tissues, representing WASF2 gene. The sequence from spermatozoa is highlighted in yellow background. Several variations detected in sperm or testis only are shown in red, and that in somatic tissues are in blue color. Note the single nucleotide variations/insertions/deletions along the sequences from different tissues with highest frequency in testis.

Figure 23. Multiple sequence alignment of GACA-tagged transcript of 523 bp originating from testis, ovary and spermatozoa, homologous to Ankyrin repeat domain-26. The sequence from spermatozoa is highlighted in yellow background. Note identical sequences in testis and spermatozoa (highlighted in red) in comparison to ovary (blue).

4.1.3.3 *Within the GATA tagged transcripts*

Next, we analyzed the GATA-tagged transcripts to look for similar sequence alterations. As discussed earlier, only 10 GATA-tagged transcripts were uncovered encompassing 4 common across the tissues and 6 restricted to testis/spermatozoa. Sequencing of 5 recombinants of each of the 6 transcripts demonstrated identical sequences both in the testis and spermatozoa. However, remaining 4 transcripts evinced several single nucleotide insertions, deletions and/or substitutions at many places. Among them was a novel transcript of 800 bp (GenBank accession numbers: EF051520 and EF050082) harboring an insertion of 18 bp at one place exclusively in spleen, and several point nucleotide changes only in the sperm and ovary in comparison to other tissues (Figure 24).

Another novel transcript of 425 bp (GenBank accession numbers: EF050083 and EF051516) also demonstrated several variations such that the point nucleotide changes were either common between the spermatozoa/gonads or spermatozoa/somatic tissues (Figure 25). Remaining novel transcripts of the 367 (Figure 26) and 282 bp (not shown) showed conserved sequences across the tissues and spermatozoa (Table 12). As already stated, the lung and heart were found to be devoid of the GATA-tagged transcripts.

4.1.4 Conservation of MASA entrapped genes across the species

To determine the evolutionary significance of all the uncovered transcripts, we studied their conservation across the species by cross-hybridization of all these genes/ gene fragments individually with the genomic DNA of 13 different species. Among the GACA-tagged transcripts, ~75% were found to be conserved faithfully across the species (>8 species) but remaining 25% were exclusive to Bovids or buffalo (Figure 27). Contrary to this, all the GATA- (Figure 28A) and 33.15- (Figure 28B) tagged transcripts were found to be cross-hybridized with genomic DNA from all the species studied but with varying signal intensities suggesting their wide conservation across the species.

Figure 24

Contd/-

Figure 24. Multiple sequence alignment of GATA-tagged novel transcript of 800 bp originating from different tissues and spermatozoa. Note the single nucleotide variations/INDELS spread throughout the sequence. The variations common to tissues are highlighted in blue color and that shared by sperm in red. Note the exclusive and major insertions of 14 bp in spleen, highlighted in blue background.

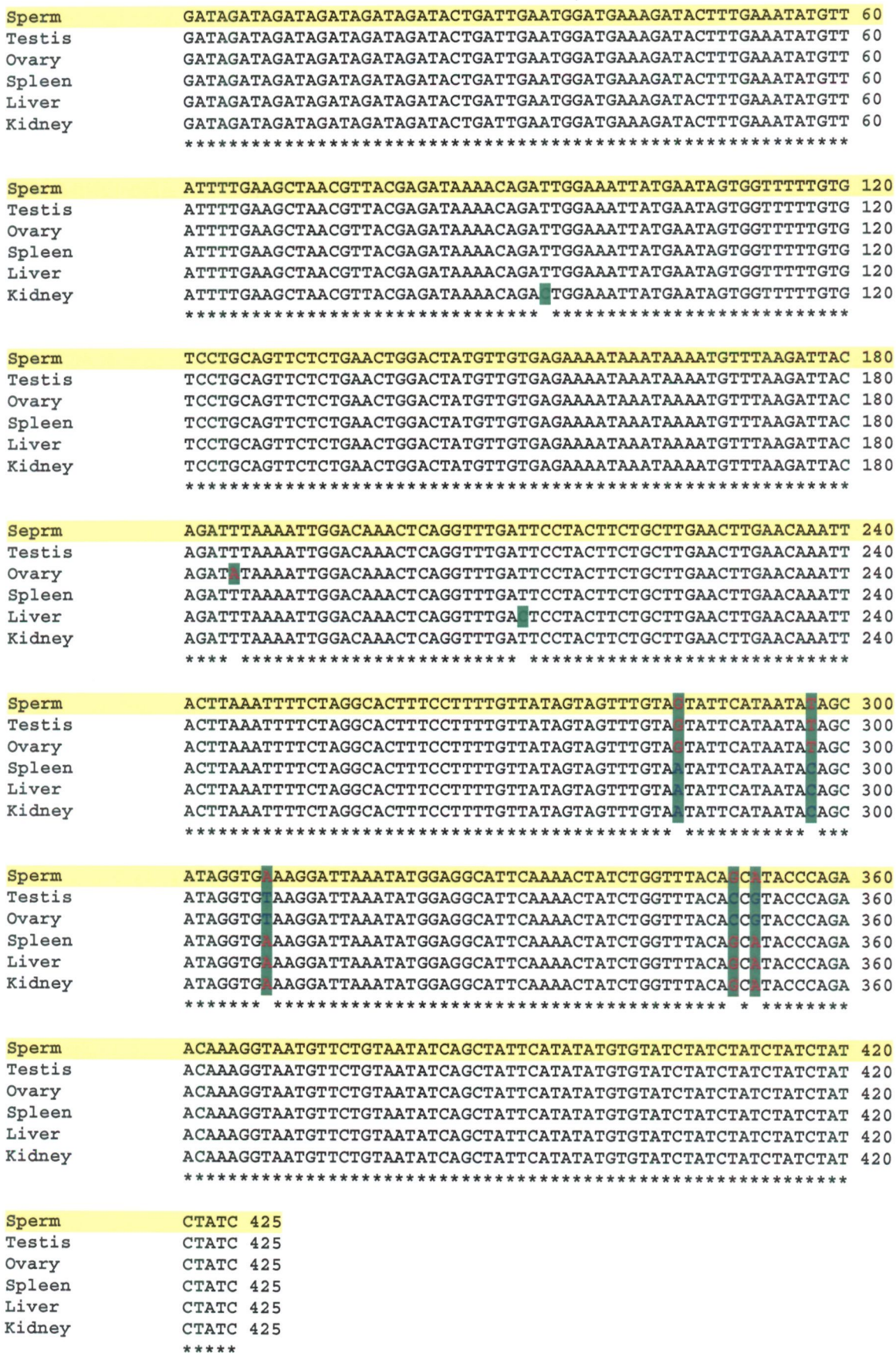


Figure 25. Multiple sequence alignment of the GATA-tagged novel transcript of 425 bp originating from different tissues and spermatozoa. The sequence from spermatozoa is highlighted in yellow background. The variations common to few tissues are highlighted in same color (blue or red).

```
SPERM       GATAGATAGATAGATAGATAGATAGCATTAAGCAAGACCCATTCTGTTGCACATATTCAA 60
Testis      GATAGATAGATAGATAGATAGATAGCATTAAGCAAGACCCATTCTGTTGCACATATTCAA 60
Kidney      GATAGATAGATAGATAGATAGATAGCATTAAGCAAGACCCATTCTGTTGCACATATTCAA 60
Ovary       GATAGATAGATAGATAGATAGATAGCATTAAGCAAGACCCATTCTGTTGCACATATTCAA 60
Liver       GATAGATAGATAGATAGATAGATAGCATTAAGCAAGACCCATTCTGTTGCACATATTCAA 60
            ************************************************************

SPERM       GCTTACATTTTTATAGGAACATTCATTTTTTGCATTGTAAACAAGAATATAATTCAGAGA 120
Testis      GCTTACATTTTTATAGGAACATTCATTTTTTGCATTGTAAACAAGAATATAATTCAGAGA 120
Kidney      GCTTACATTTTTATAGGAACATTCATTTTTTGCATTGTAAACAAGAATATAATTCAGAGA 120
Ovary       GCTTACATTTTTATAGGAACATTCATTTTTTGCATTGTAAACAAGAATATAATTCAGAGA 120
Liver       GCTTACATTTTTATAGGAACATTCATTTTTTGCATTGTAAACAAGAATATAATTCAGAGA 120
            ************************************************************

SPERM       AGAAATTACAAATTCTATATTTTGTTGGTATTTTATTTATTGCATCTGACCTTTAAGAAC 180
Testis      AGAAATTACAAATTCTATATTTTGTTGGTATTTTATTTATTGCATCTGACCTTTAAGAAC 180
Kidney      AGAAATTACAAATTCTATATTTTGTTGGTATTTTATTTATTGCATCTGACCTTTAAGAAC 180
Ovary       AGAAATTACAAATTCTATATTTTGTTGGTATTTTATTTATTGCATCTGACCTTTAAGAAC 180
Liver       AGAAATTACAAATTCTATATTTTGTTGGTATTTTATTTATTGCATCTGACCTTTAAGAAC 180
            ************************************************************

SPERM       AGAATTTATGTAAAACCAAAGACATGTTTCCTGTAATTACAGCAAAATGATAATGATTAT 240
Testis      AGAATTTATGTAAAACCAAAGACATGTTTCCTGTAATTACAGCAAAATGATAATGATTAT 240
Kidney      AGAATTTATGTAAAACCAAAGACATGTTTCCTGTAATTACAGCAAAATGATAATGATTAT 240
Ovary       AGAATTTATGTAAAACCAAAGACATGTTTCCTGTAATTACAGCAAAATGATAATGATTAT 240
Liver       AGAATTTATGTAAAACCAAAGACATGTTTCCTGTAATTACAGCAAAATGATAATGATTAT 240
            ************************************************************

SPERM       CACAAACAATCTCTAAATATTAGTGTCTATAGATGATTGAAAGTGAAAGTACATTGTAGG 300
Testis      CACAAACAATCTCTAAATATTAGTGTCTATAGATGATTGAAAGTGAAAGTACATTGTAGG 300
Kidney      CACAAACAATCTCTAAATATTAGTGTCTATAGATGATTGAAAGTGAAAGTACATTGTAGG 300
Ovary       CACAAACAATCTCTAAATATTAGTGTCTATAGATGATTGAAAGTGAAAGTACATTGTAGG 300
Liver       CACAAACAATCTCTAAATATTAGTGTCTATAGATGATTGAAAGTGAAAGTACATTGTAGG 300
            ************************************************************

SPERM       TGGAGTCTTTACCAACTGAACTATCAGGAAAGCCCAGTGACTATATCTATCTATCTATCT 360
Testis      TGGAGTCTTTACCAACTGAACTATCAGGACAGCCCAGTGACTATATCTATCTATCTATCT 360
Kidney      TGGAGTCTTTACCAACTGAACTATCAGGAAAGCCCAGTGACTATATCTATCTATCTATCT 360
Ovary       TGGAGTCTTTACCAACTGAACTATCAGGAAAGCCCAGTGACTATATCTATCTATCTATCT 360
Liver       TGGAGTCTTTACCAACTGAACTATCAGGAAAGCCCAGTGACTATATCTATCTATCTATCT 360
            ***************************** ******************************

SPERM       ATCTATC 367
Testis      ATCTATC 367
Kidney      ATCTATC 367
Ovary       ATCTATC 367
Liver       ATCTATC 367
            *******
```

Figure 26. Multiple sequence alignment of the GATA-tagged novel transcript of 367 bp originating from different tissues and spermatozoa. Note this fragment showed almost identical sequences except for few random point nucleotide changes.

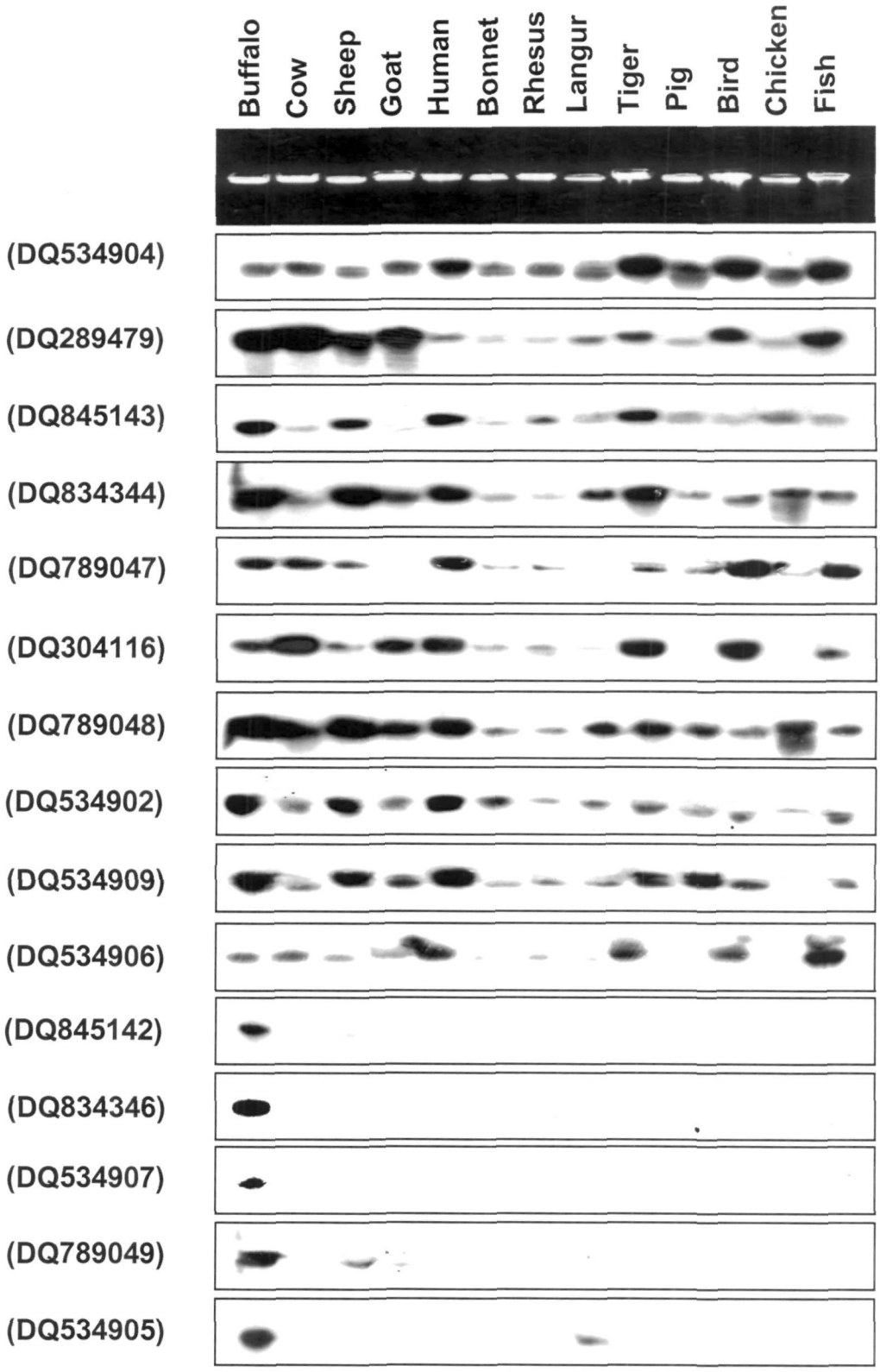


Figure 27. Cross-hybridization of genomic DNA from different species with the recombinant clones for the GACA uncovered gene fragments. The names of the species are given on the top, and the autoradiograms for the respective gene/gene fragments on the left.

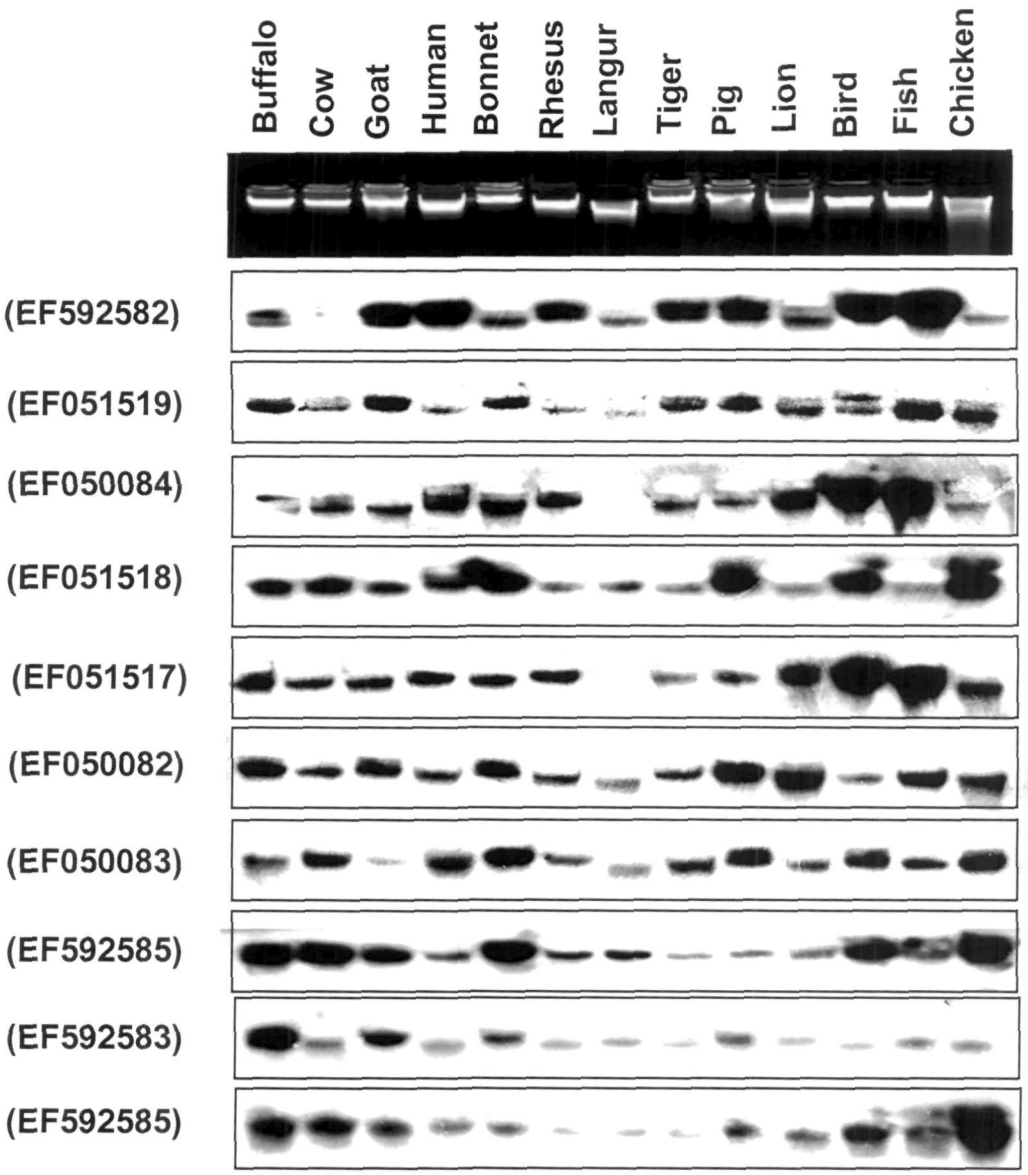


Figure 28A. Cross-hybridization of genomic DNA from different species with recombinant clones for GATA uncovered gene fragments. The names of the species are given on the top, and the autoradiograms for the respective gene/gene fragments on the left.

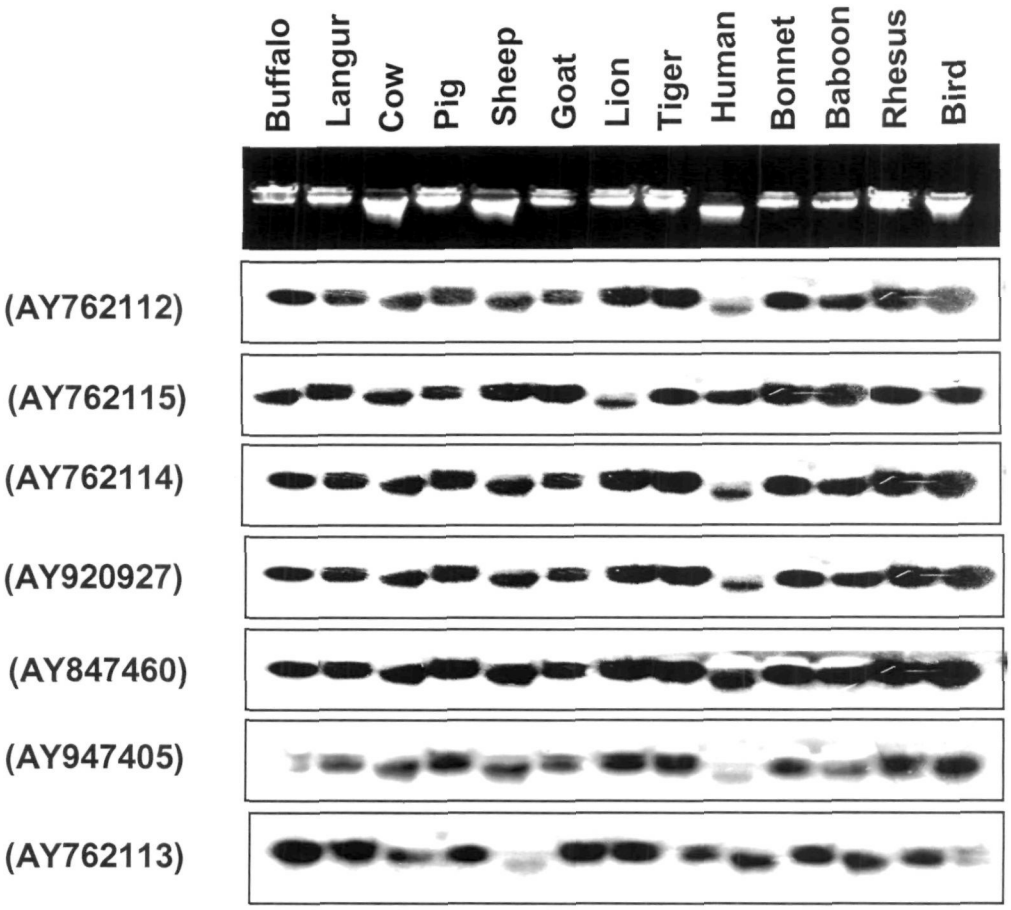


Figure 28B. Cross-hybridization of genomic DNA from different species with recombinant clones for 33.15 uncovered gene fragments. The names of the species are given on the top panel, and the autoradiograms for the gene/gene fragments are given on the left.

4.1.5 Copy number status of the uncovered genes

Following the sequence analyses and evolutionary studies, the copy number of these repeat tagged gene/gene fragments was calculated by extrapolation of the straight curve obtained in a Real Time PCR using 10 fold dilution series of the respective recombinant plasmids. Extrapolation of these standard curves demonstrated the copy number status of the genes identified with the consensus of 33.15 repeat (Figure 29) and simple quadruplets of GACA (Figure 30) and GATA (Figure 31) which varied from 1 to 65 per haploid genome in buffalo. Here, representative Real Time plots, standard curves and dissociation curves have been shown to demonstrate the copy number calculation for these repeat tagged genes.

The copy number for all the 33.15 repeat tagged genes were calculated as 1-2 per copies per haploid genome in the water buffalo (Table 13). Out of the 32 GACA-tagged transcripts studied, nineteen had single copy; eleven, 2-3; one each with 8-13 and 25-65 copies (Table 14). Similarly, of the 8 GATA-tagged transcripts, three were single copy and five had 2-5 copies (Table 14). Briefly, the copy number of the GACA- and GATA-tagged genes varied from 1 for 50%, 2-5 for 45% and 8-65 for the remaining 5% for all the GACA/GATA tagged genes/gene fragments (Table 14).

4.1.6 Differential expression of the repeat tagged genes

After ascertaining the tissue-specific organizational variation, their comparative expression profiles were studied to explore their functional status in different somatic tissues, gonads and spermatozoa. The expression study for each MASA uncovered gene/gene fragments was done first by RNA slot blot hybridization and RT-PCR analysis (for e.g. Figures 14 and 15). The quantitative expressional analysis was then performed for individual transcript using β-actin as an internal control in Real Time PCR.

In the Real Time PCR analyses, primer specificity for the uncovered genes was established using five fold dilution series of the template cDNA (Figure 32A). Straight standard curve with a slope = (-3.4) to (-3.6) and a

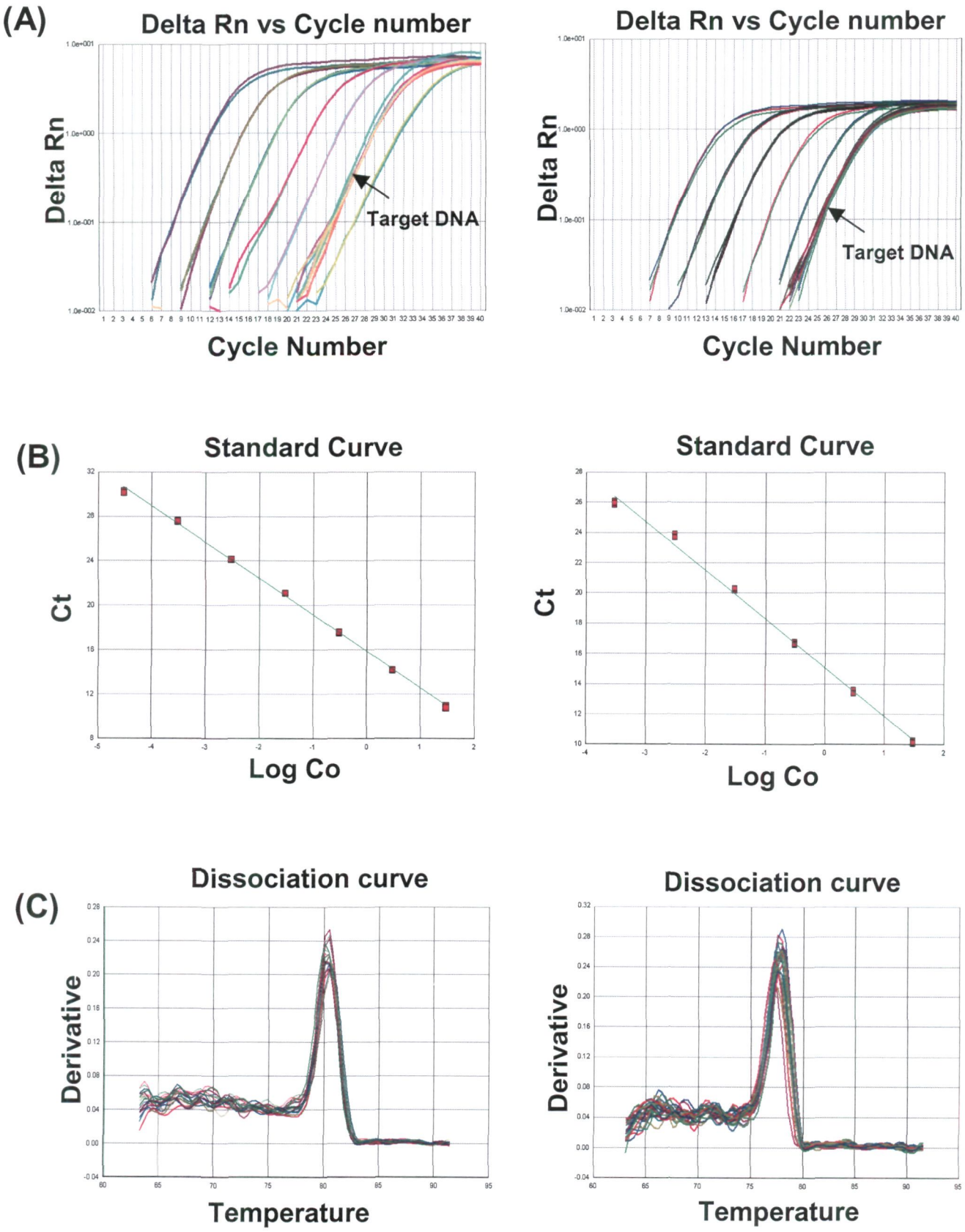


Figure 29. Representative Real Time PCR amplification plots for copy number calculation using the target DNA and 10 fold dilution series (From 300 million to 300 copies) of recombinant plasmids containing 33.15 uncovered genes **(A).** The assays were performed using SYBR green chemistry and the derived Standard **(B)** and Dissociation Curves **(C)** are also shown. Single peak in the dissociation curve shows high specificity of the primers. Ct and Log Co denote the cycle threshold and log of concentration, respectively.

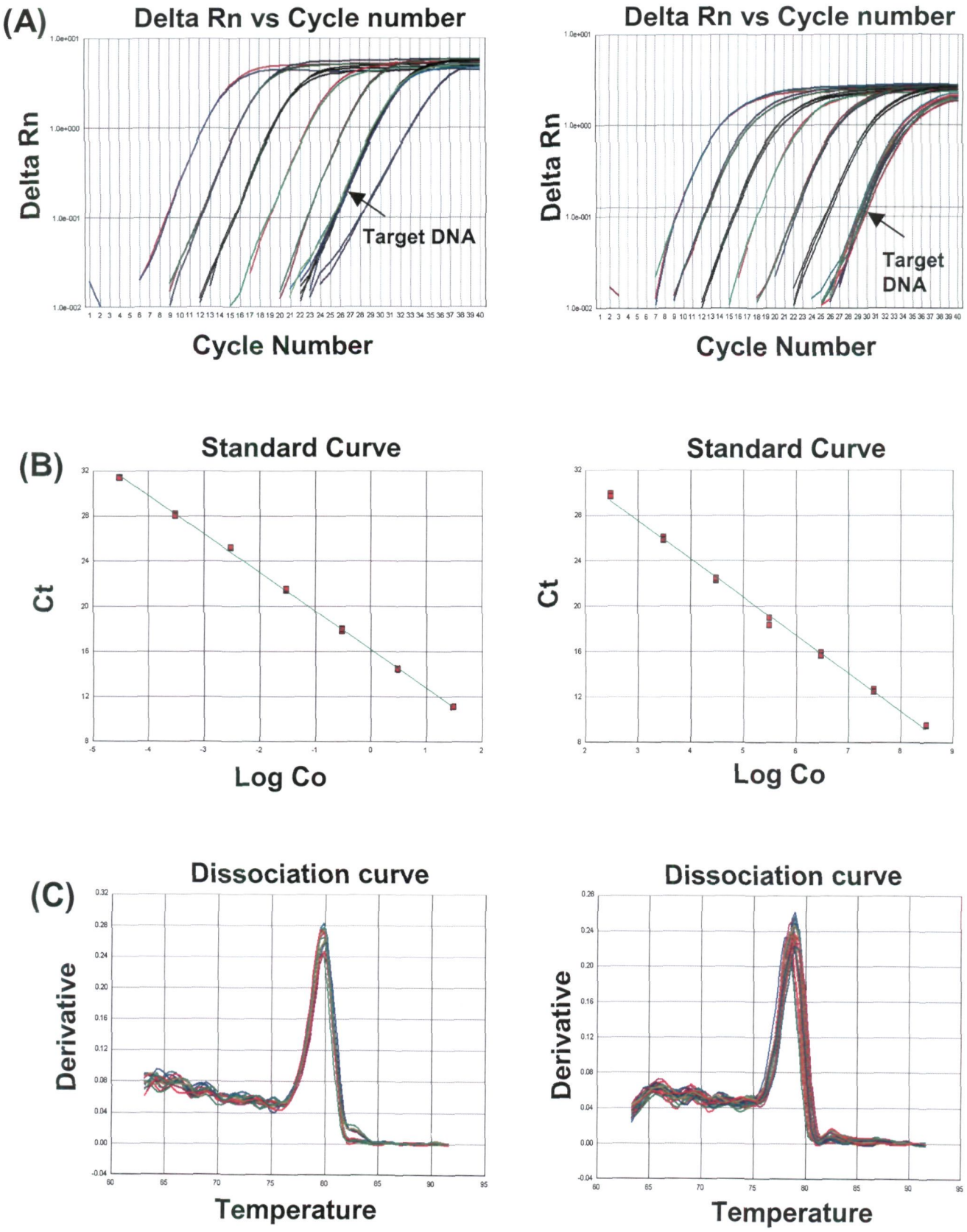


Figure 30. Representative Real Time PCR amplification plots for copy number calculation using target genomic DNA and 10 fold dilution series (From 300 million to 300 copies) of recombinant plasmids containing GACA uncovered genes **(A).** The assays were performed using SYBR green chemistry and the derived Standard **(B)** and Dissociation Curves **(C)** are also shown. Single peak in the dissociation curve shows high specificity of the primer. Ct and Log Co denote the cycle threshold and log of concentration, respectively.

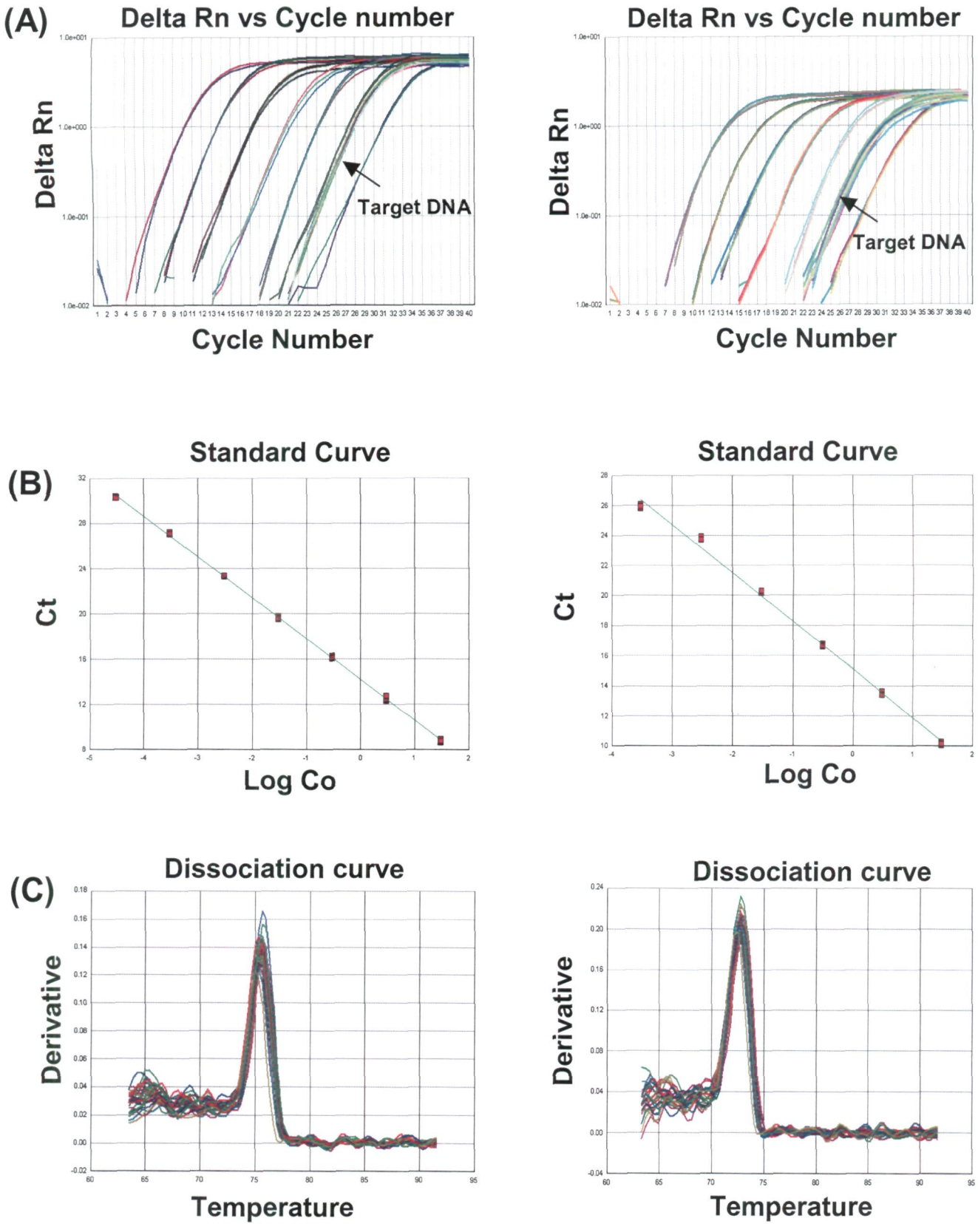


Figure 31. Representative Real Time PCR amplification plots for copy number calculation using the target genomic DNA and 10 fold dilution series (From 300 million to 300 copies) of recombinant plasmids containing GATA uncovered genes **(A).** The derived Standard **(B)** and Dissociation Curves **(C)** using SYBR Green Dye. Single peak in the dissociation curve shows high specificity of the primer.

Table 13: Relative quantitative expression and Copy number status of the genes/gene fragments tagged with the consensus of 33.15 repeat loci#

S.No.	Clone ID	Gene Accession Numbers	Relative expression in different tissues (*in folds*)							Relative expression in *spermatozoa* from four buffaloes				Copy number status per haploid genome
			Testis	Ovary	Spleen	Liver	Lung	Kidney	Heart	SP1	SP2	SP3	SP4	
1.	pJC7	AY847460	65	20	50	35	*Cb*	15	5	64	39	52	47	1-2 copies
2.	pJC5	AY762114	294	30	83	23	18	8	*1*	388	238	416	147	
3.	pJC4	AY762112	45	12	23	35	5	*Cb*	15	39	21	34	23	
4.	pJC6	AY762115	3	2	2	3	*2*	*Cb*	*Cb*	2	3	2	2	
5.	pJC8	AY920927	175	*28*	44	51	16	*Cb*	2	74	49	111	84	
6.	pJC10	AY947405	8	15	3	171	*Cb*	4	*10*	6	6	8	5	
7.	pJSC44	EU348484	14	14	3	6	8	3	*Cb*	111	239	239	104	
8.	pJSC45	EU348485	1	3	3	6	*3*	1	*Cb*	13	25	30	32	
9.	pJSC46	EU348486	104	97	37	119	128	*42*	*Cb*	955	588	1910	724	
10.	pJSC50	EU348490	16	5	16	9	21	2	*Cb*	137	64	119	42	
11.	pJSC47	EU348487	21	1	5	3	3	*Cb*	3	59	39	69	29	
12.	pJSC49	EU348489	84	9	11	7	*5*	*Cb*	6	97	64	157	51	
13.	pJSC48	EU348488	74	18	7	6	15	2	*Cb*	274	168	315	128	
14.	pJSC43	EU348483	97	16	6	8	*8*	*Cb*	3	84	64	97	32	
15.	pJSC42	EU348482	1097	147	104	97	*79*	37	*Cb*	1351	832	1448	776	
16.	pJSC41	EU348481	9	*56*	15	6	9	2	*Cb*	119	55	104	45	
17.	pJSC40	EU348480	14	9	26	9	11	*Cb*	26	23	18	14	49	
18.	pJSC39	EU348479	24	18	8	4	6	*Cb*	3	158	111	158	137	

Table 14: Relative quantitative expression and Copy number status of the genes/gene fragments tagged with GACA & GATA repeat motifs

S.No.	Clone ID	Gene Accession Numbers	Relative expression in different tissues (in folds)							Relative expression in spermatozoa from four buffaloes				Copy number status per haploid genome	
			Testis	Ovary	Spleen	Liver	Lung	Kidney	Heart	SP1	SP2	SP3	SP4		
A. For transcripts tagged with GACA repeat motif														**In blood**	**In germline**
1.	pJC40	DQ494483	194	21	23	17	2	*Cb*	3	274	181	147	239	1	1
2.	pJC42	DQ494485	32	8	2	30	7	51	*Cb*	29	17	21	27	2-3	2-3
3.	pJC52	DQ534910	512	32	34	83	24	60	*Cb*	6	9	5	7	3	3
4.	pJC54	DQ834345	208	28	15	69	3	*Cb*	1	107	119	97	157	1	1
5.	pJC29	DQ289479	15	10	13	45	*Cb*	20	22	49	52	32	45	1	1
6.	pJC35	DQ304116	147	24	51	45	3	*Cb*	3	39	32	51	39	1-2	1-2
7.	pJC44	DQ534902	44	21	17	34	11	25	*Cb*	97	111	97	128	1	1
8.	pJC46	DQ534904	7	6	2	18	3	*Cb*	3	14	22	11	12	2	2
9.	pJC47	DQ534905	34	11	14	91	*Cb*	14	2	73	97	87	84	1	1
10.	pJC49	DQ534907	3521	891	330	637	238	*Cb*	630	2896	4792	2702	3326	1	1
11.	pJC51	DQ534909	1663	157	338	2521	3	5	*Cb*	362	239	512	676	1	1
12.	pJC53	DQ834344	17	13	5	29	4	*Cb*	45	18	14	12	16	2	2
13.	pJSC11	DQ845144	4390	1176	664	1097	*Cb*	2	1195	6616	5120	8526	7342	25-65	30-65
14.	pJSC1	DQ789045	46	35	40	36	12	15	*Cb*	36	21	23	27	1	1
16.	pJSC3	DQ789047	156	45	12	87	*Cb*	2	37	1176	724	776	1440	1	1
17.	pJSC4	DQ789048	149	222	376	34	*Cb*	10	6	675	630	608	588	2	2
18.	pJSC5	DQ789049	128	2	2	9	2	*Cb*	2	62	47	41	38	2	2
19.	pJSC6	DQ834346	31	21	30	14	*Cb*	13	51	52	97	84	55	1	1
20.	pJSC9	DQ845142	53	4	3	4	3	*Cb*	3	15	14	15	14	2	2
21.	pJSC10	DQ845143	3	3	12	4	*Cb*	14	3	6	4	9	3	3	3
22.	pJSC12	DQ845145	91	6	28	52	2	6	*Cb*	138	97	119	97	1	1
23.	pJSC13	DQ845146	228	181	246	34	2	74	*Cb*	1782	1910	1097	1351	1	1

24.	pJSC15	DQ904037	39	4	26	13	*Cb*	5	2	49	35	45	39	8-13	8-10
25.	pJSC16	DQ904038	31	*Cb*	14	9	2	2	1	117	112	127	118	1	1
26.	pJSC17	DQ904039	27	22	19	15	9	16	*Cb*	14	29	18	20	1	1
27.	pJSC18	DQ913640	18	7	42	28	*Cb*	9	2	34	23	42	23	1	1
28.	pJSC19	DQ913641	85	24	35	28	2	13	*Cb*	69	68	83	52	2	2
29.	pJSC20	DQ913642	89	74	88	81	*Cb*	65	74	81	88	71	82	1	1
30.	pJSC22	DQ913644	*Cb*	4	4	2	2	8	3	75	69	54	61	2-3	2
31.	pJSC23	DQ913645	2	2	12	4	10	2	*Cb*	55	73	41	67	1	1
32.	pJSC24	DQ913646	2	1	12	2	*Cb*	2	1	48	27	42	32	1	1
33.	pJSC25	DQ916743	149	127	104	21	109	64	*Cb*	239	194	195	256	1	1
B. For transcripts tagged with GATA repeat motif															
1.	pJSC28	EF050082	114	58	16	5	*Cb*	2	3	51	48	34	42	2-4	2-4
2.	pJSC30	EF050084	169	20	65	48	*Cb*	1	1	168	128	113	137	1	1
3.	pJSC31	EF051516	65	30	23	35	*Cb*	10	5	59	48	53	43	2	2
4.	pJSC32	EF051517	326	33	28	52	*Cb*	8	3	1351	1261	1351	1261	1	1
5.	pJSC33	EF051518	239	57	14	68	2	44	*Cb*	42	68	55	73	3-5	3-5
6.	pJSC34	EF051519	490	78	19	14	*Cb*	37	3	589	510	465	610	2	2
7.	pJC87	EF592582	386	39	14	9	*Cb*	2	3	314	296	357	260	1	1
8.	pJC88	EF592583	134	87	93	102	4	15	*Cb*	201	174	124	145	1-2	1-2

The expression for gene fragments tagged with GACA repeat is described in **(A)** whereas for GATA-tagged ones in **(B)**. Note the highest expression of most of the GACA-tagged and all GATA-tagged genes in testis and/or spermatozoa.

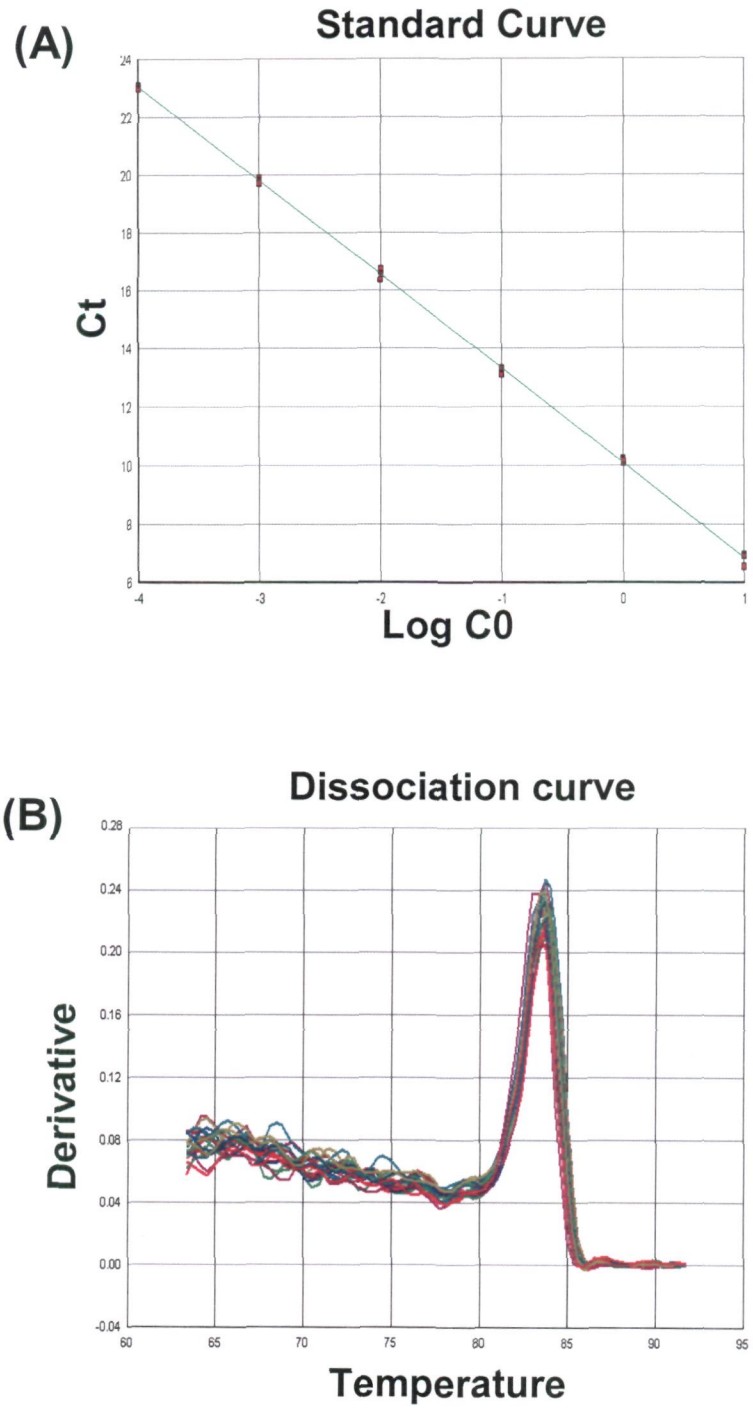


Figure 32. Standard **(A)** and Dissociation Curves **(B)** based on Real Time PCR using SYBR Green Dye and five fold dilution series of the cDNA samples. Single peak in the dissociation curve shows high specificity of the primers.

single peak in the dissociation protocol established the specificity of the primers (Figure 32B). For the accuracy, the expression study was conducted using three different dilutions of cDNA. The results so obtained were substantiated further by expression data from five additional animals.

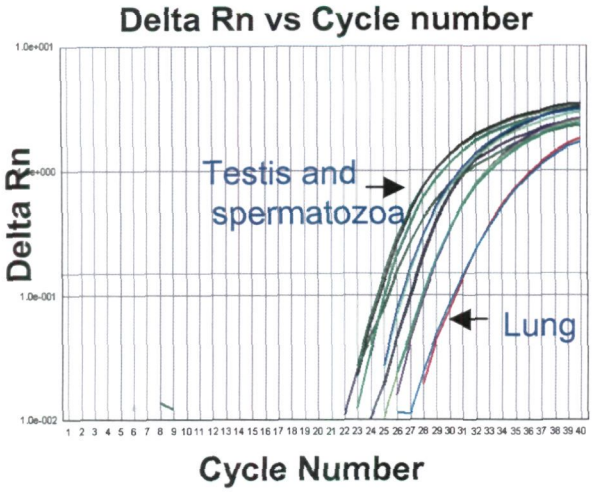
4.1.6.1 *Of the 33.15 repeat tagged genes*

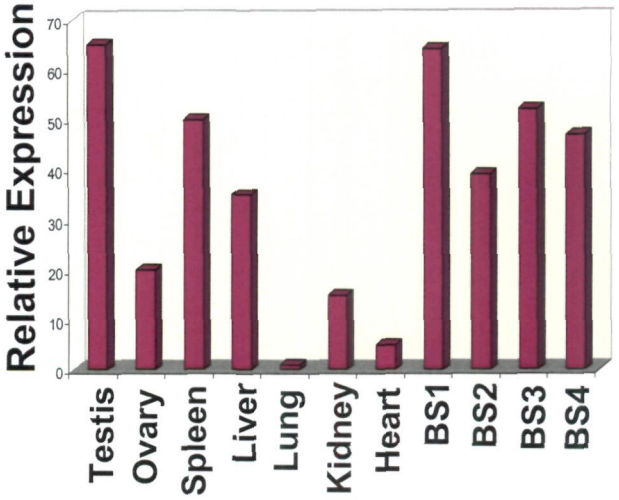
Of the 7 gene fragments, 5 showed positive signals in all the tissues studied whereas 2 remained restricted to a particular tissue (Figure 14A). The 846 bp and 487 bp fragments also showed almost uniform signals in all the tissues (Figure 14B, panel a-b). The 324 bp fragment resulted in strong signals in testis, ovary and spleen but very faint ones in liver, heart and lung with the no signal detected in kidney (Figure 14B, panel c). However, β-actin used as positive control showed similar signal intensity in all the lanes (Figure 14B, panels d and g). The 576 bp fragment again showed strong signal in spleen and a faint one in liver (Figure 14B, panel e) whereas 602 bp one uncovered signals in all the tissues (Figure 14B, panel f). The 1263 bp fragment having homology to *Smoc*-1 gene showed exclusive signal in the liver (shown in the section 4.2.2). However, upon long exposure, *Smoc*-1 also showed its faint signal in the testis and ovary besides that prominent in the liver.

By the Real Time PCR, relative expression of the tissue-originated genes, *Smoc*-1, *AKL*, *TCRL*, *LRRN6A* and *TCRGL*, was studied. *Smoc*-1 gene showed highest expression (160 folds) in liver (Figure 33A) and Adenylate kinase like (AKL) gene (65 folds) in testis (Figure 33B), compared to that in lung as endogenous control. TCRL gene showed highest expression (83 folds) in spleen relatively to that in heart (Figure 33C). The highest expression of LRRN6A (42 folds) and TCRGL (165 folds) genes was detected in testis compared to that in kidney as endogenous control (Figure 33D-E).

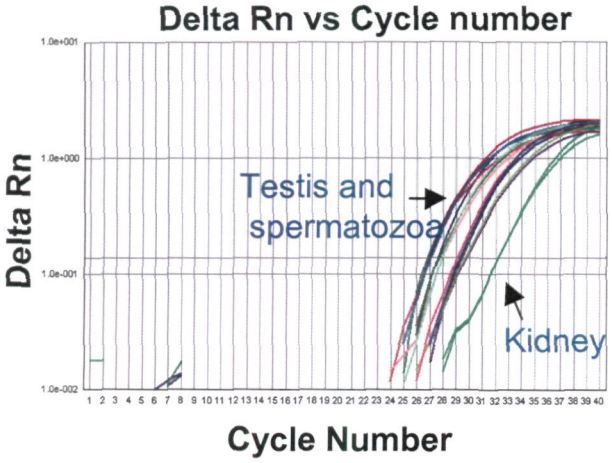
Interestingly, the relative expressional studies of all the spermatozoal transcripts revealed that out of 12, only one transcript (GenBank expression number: EU348480) was showing uniform expression in all the tissues, 5 transcripts showed highest expression in

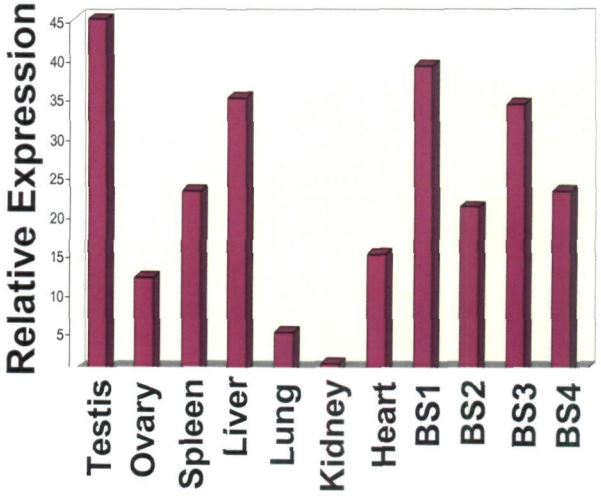
(A. pJC7)



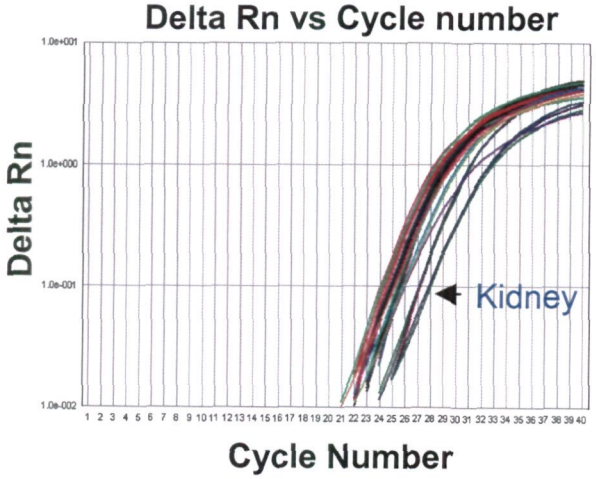


(B. pJC4)





(C. pJC6)



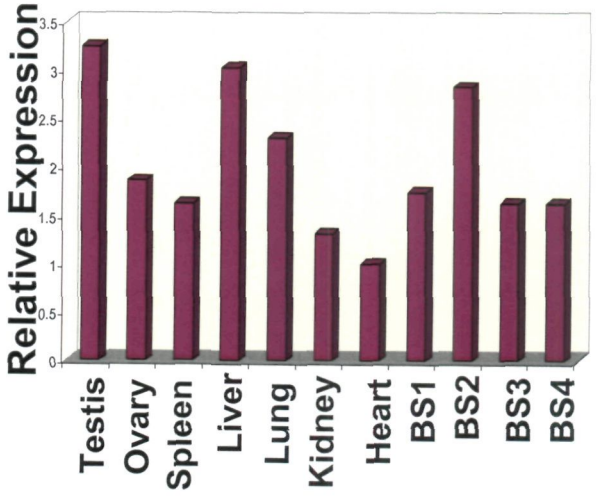


Figure 33

Contd/-

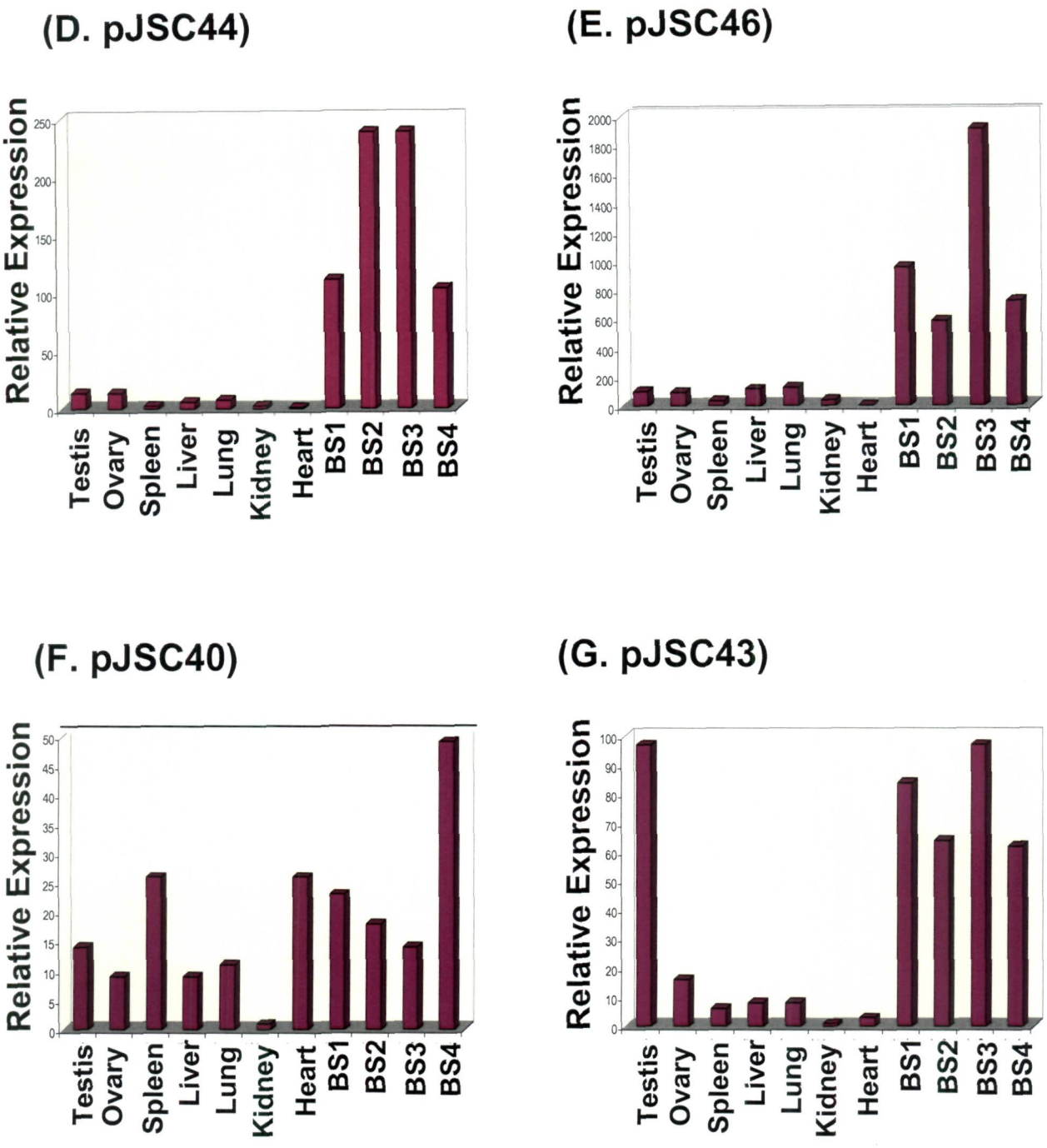


Figure 33. Real Time PCR amplification plots showing relative expression of the 33.15 tagged genes and the respective bar diagrams demonstrating comparative analysis of the expression across the tissues and spermatozoa of buffalo. The expression plots and respective bar diagrams for somatic originated transcripts **(A-C)**, and for spermatozoa originated ones **(D-G)** are shown. Note the highest expression of most of the fragments in the testis and spermatozoa.

the spermatozoa and 6 in both the testis and spermatozoa compared to that in somatic tissues and ovary (Table 13)

4.1.6.2 *Of the GACA tagged transcripts*

A total of 32 GACA-tagged transcripts were studied (Table 14) for the quantitative expressional studies using Real Time PCR assays. When the expression was compared between somatic tissues and gonads; ~50% transcripts evidenced highest expression in testis, ~20% in spleen/liver, and remaining ~30% with uniform expression in all the tissues. Further, the comparative expression of these transcripts amongst tissues and spermatozoa unveiled surprising observations. First of all, 14 transcripts showed highest expression in the spermatozoa followed by in testis, and 3 remained exclusive to the spermatozoa. Secondly, 2 transcripts demonstrated unique expression in testis, 4 in liver/spleen and 9 with consistent expression in all the sources studied.

Among the uncovered transcripts, the highest expression observed was of Ankyrin repeat domain (3400-4390 folds in testis and 5120-8526 folds in spermatozoa), followed by the WASF2 gene (3521 folds in testis and 2896 to 4792 folds in spermatozoa) (Figure 34A). The testis-specific expression was observed for only 2 transcripts namely Ubiquitin-associated protein-1 (Ubap1) (Accession no. DQ534910), and 1.1 kb transcript representing ß-transducin repeat (Accession no. DQ304116) (Figure 34B). Ubap1 and ß-transducin repeat showed 150-200 and 100-160 folds expression respectively in testis compared to that in kidney as calibrator. Some transcripts such as non-POU domain containing, octamer-binding gene (Accession no. DQ789047) showed either highest or exclusive expression in the spermatozoa (Figure 34C), whereas other transcripts for eg. HBGF-1 (Accession no. DQ534904), potassium voltage gated channel, member C (Accession no. DQ904039), etc. demonstrated uniform expression in all the tissues (Figure 34D-E).

4.1.6.3 *Of the GATA tagged transcripts*

Following this approach, we pursued with the expressional analysis of all the 10 transcripts uncovered by GATA repeat. Strikingly, all of them

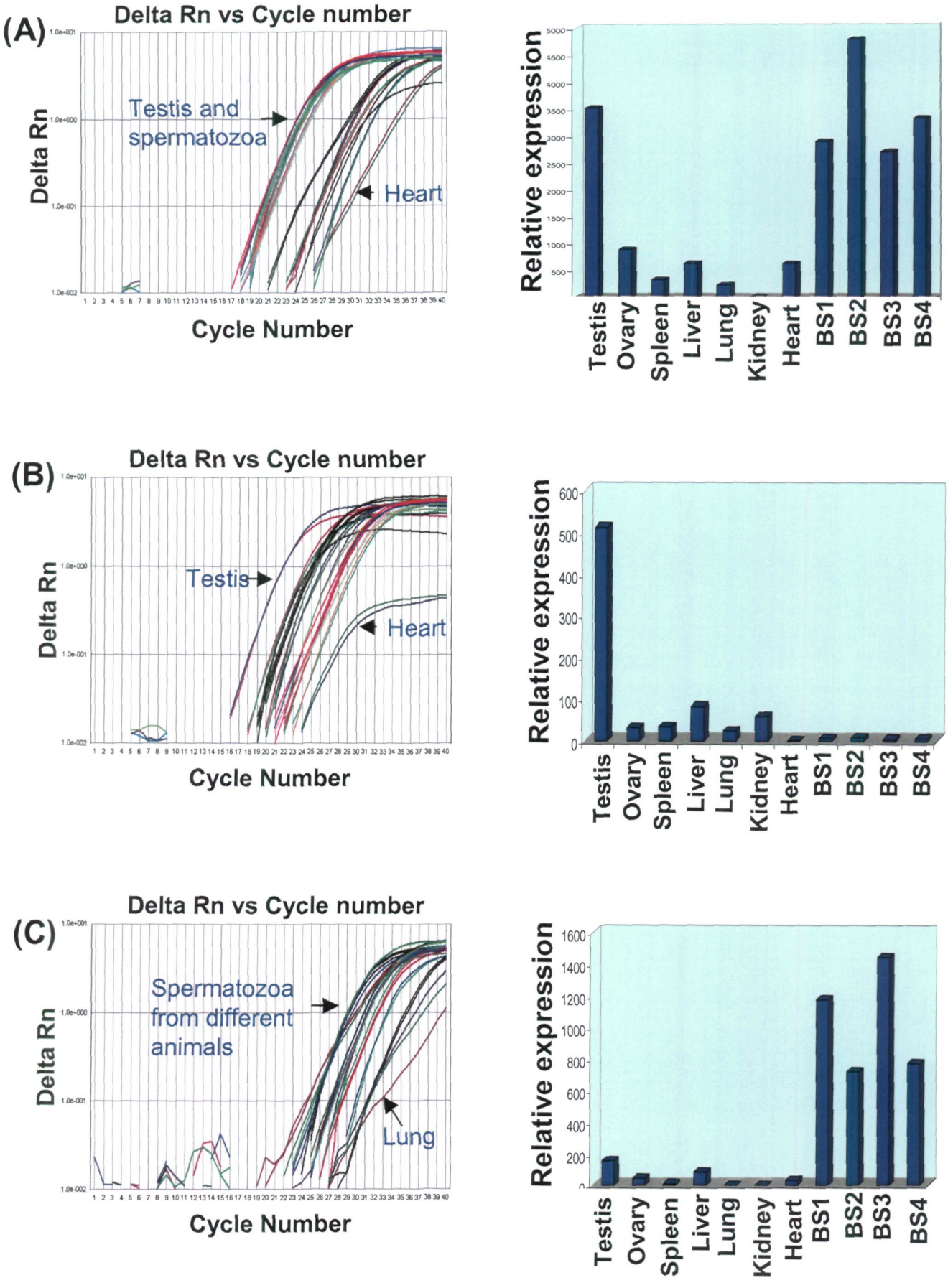


Figure 34

Contd/-

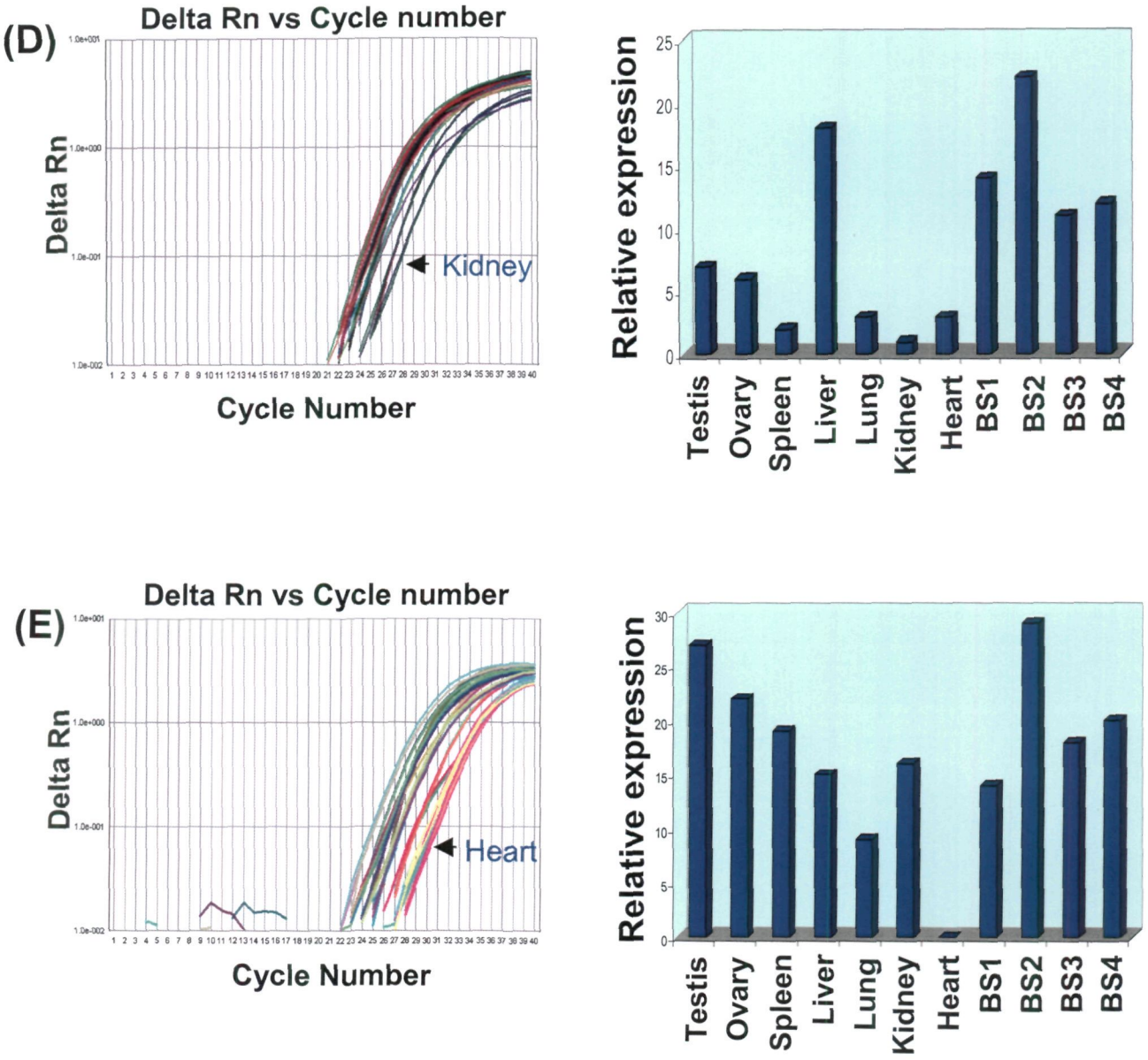


Figure 34. Real Time PCR amplification plots showing quantitative expression of representative GACA-tagged transcripts demonstrating their expressional variations among somatic/gonadal tissues and spermatozoa. Four types of expressional profiles were uncovered with GACA; some transcripts with highest expression in testis and spermatozoa e.g. Ankyrin repeat domain **(A)**, few in testis only e.g. Ubap1 **(B)**, few in spermatozoa only e.g. novel pJSC3 **(C)**, and others distributed almost uniformly in all the tissues e.g. HBGF-1 **(D)**. For details, see table 14 and text.

demonstrated highest or unique expression either in testis or spermatozoa or both, compared to that in other somatic tissues (For details, please see table 14 & figure 35A-D). Lung and heart both showed almost negligible expression which substantiated the absence of the GATA-tagged transcripts in these somatic tissues.

Thus, most of the 33.15-, GACA- and all the GATA-tagged transcripts were found to show exclusive or highest expression in the testis and/or spermatozoa. Detailed expressional analysis of all the GACA/GATA tagged transcripts including their clone IDs and accession numbers has been given in the table 14.

4.1.7 Chromosomal mapping

Chromosomal mapping was done for two candidate genes that have been characterized in other species and showed homology along their entire lengths. These two fragments of 523 and 217 bp represented the genes, Ankyrin repeat domain-26 (ANKD26) and Ubiquitin associated protein 1 (Ubap1), respectively. We performed chromosomal localization for these genes following signal amplification based method by Fluorescent *in situ* hybridization (FISH), as described earlier in the section 3.11. The Ubap1 gene was mapped onto short arm of the metacentric chromosome 3 (Figure 36) whereas Ankyrin repeat domain-26 (ANKD26) onto the proximal end of short arm of the sub-metacentric chromosome 4 in water buffalo (Figure 37).

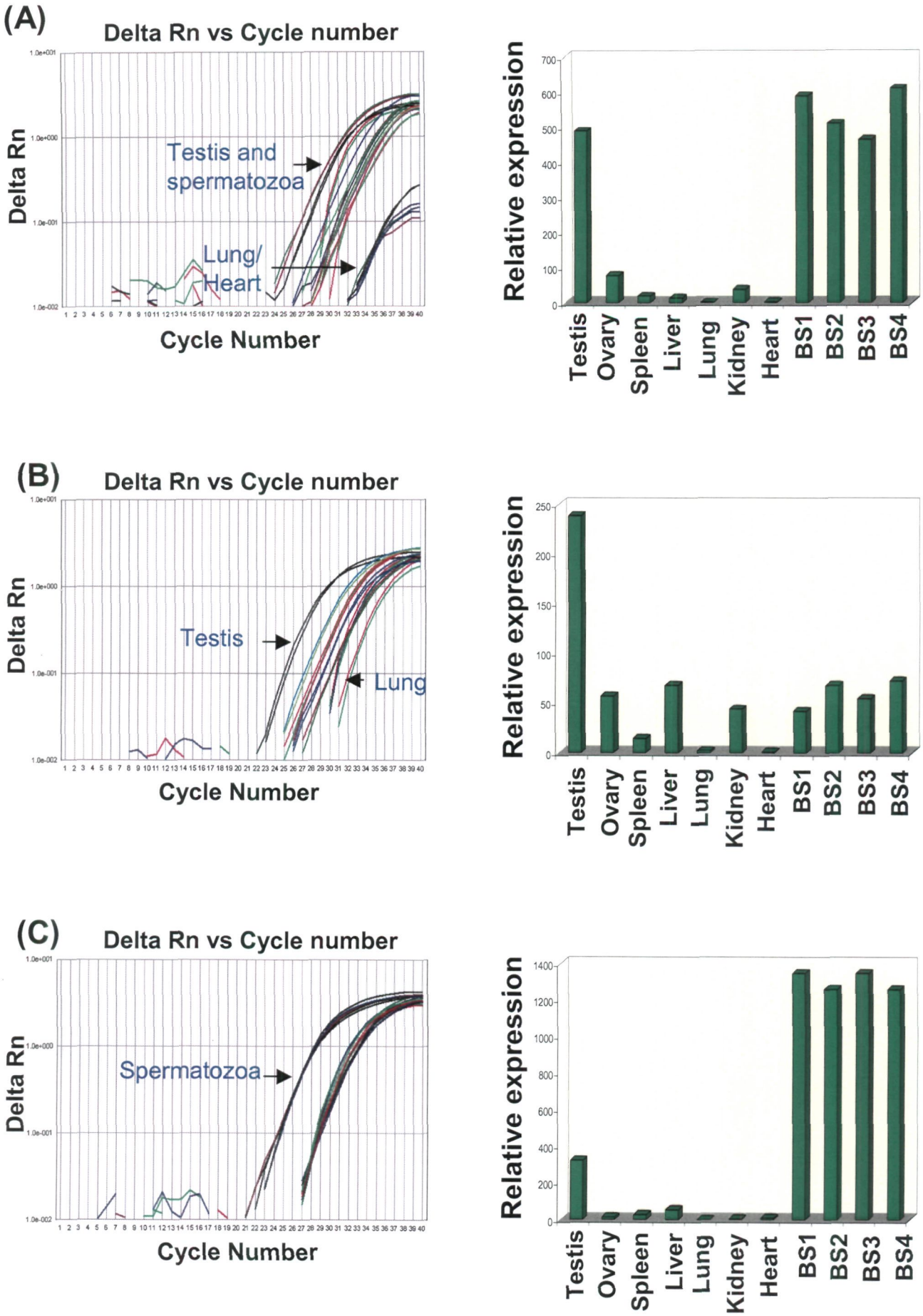

(A)
Delta Rn vs Cycle number
Delta Rn
Testis and spermatozoa
Lung/ Heart
Cycle Number
Relative expression
Testis
Ovary
Spleen
Liver
Lung
Kidney
Heart
BS1
BS2
BS3
BS4
(B)
Delta Rn vs Cycle number
Delta Rn
Testis
Lung
Cycle Number
Relative expression
Testis
Ovary
Spleen
Liver
Lung
Kidney
Heart
BS1
BS2
BS3
BS4
(C)
Delta Rn vs Cycle number
Delta Rn
Spermatozoa
Cycle Number
Relative expression
Testis
Ovary
Spleen
Liver
Lung
Kidney
Heart
BS1
BS2
BS3
BS4

Figure 35

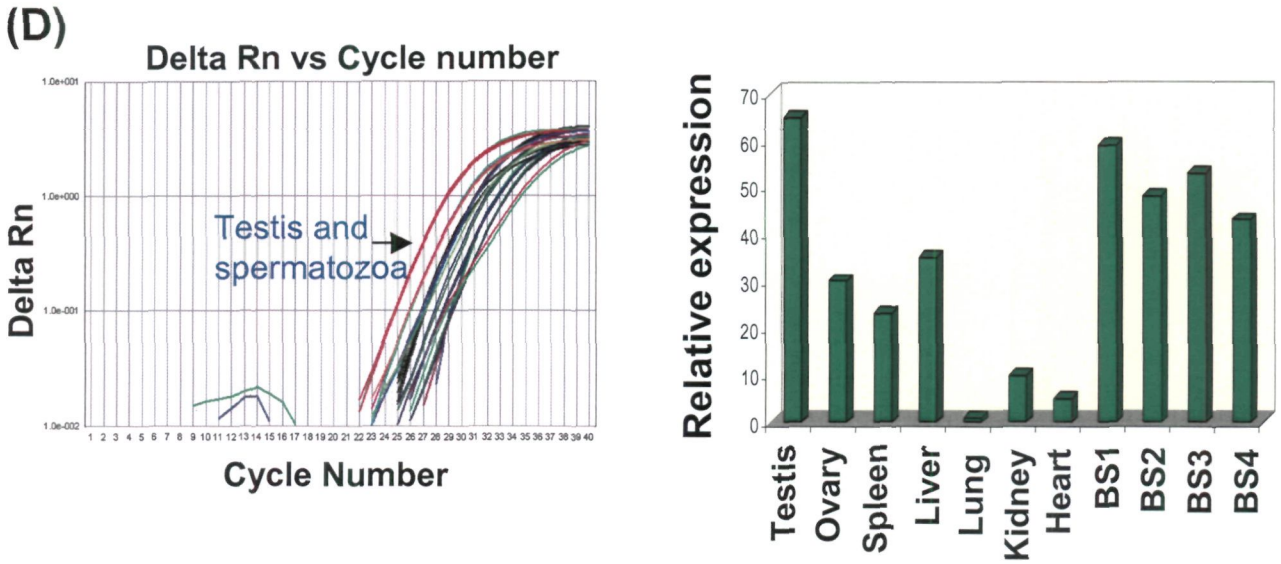
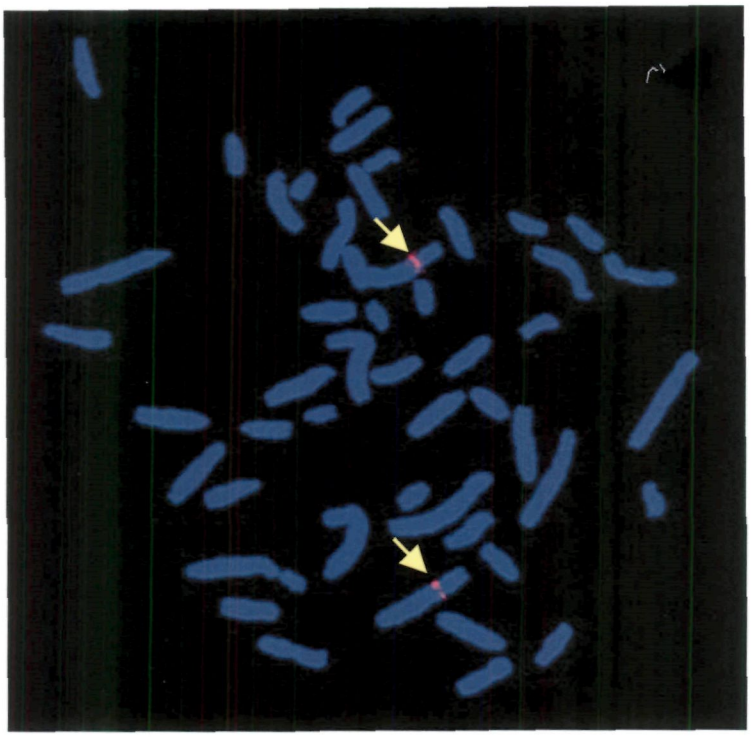
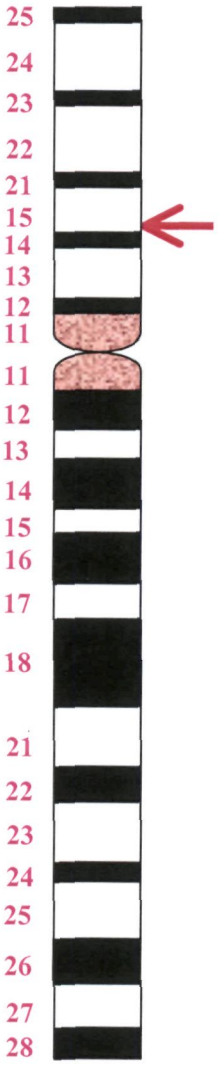


Figure 35. Real Time PCR amplification plots showing quantitative expression of representative GATA-tagged transcripts demonstrating their expressional variations among somatic/gonadal tissues and spermatozoa. Three types of expressional profiles were observed for GATA-tagged transcripts; some showed highest expression both in testis and spermatozoa e.g. novel pJSC34 **(A)**, few in testis only e.g. novel pJSC33 **(B)**, few others in spermatozoa only e.g. novel pJSC32 **(C)**, and others highest in testis and spermatozoa but with minimal variation in comparison to somatic tissues e.g. novel pJSC31 **(D)**. For details, see table 14 and text.

(A)



(B)



Chromosome 3

Figure 36. Chromosomal mapping for the candidate Ubap1 gene onto the short arm of metacentric chromosome 3 **(A)** and detailed mapping for the same with respect to its position on the G-banded ideogram following ISCNDB 2000 has been shown in the figure **(B)**.

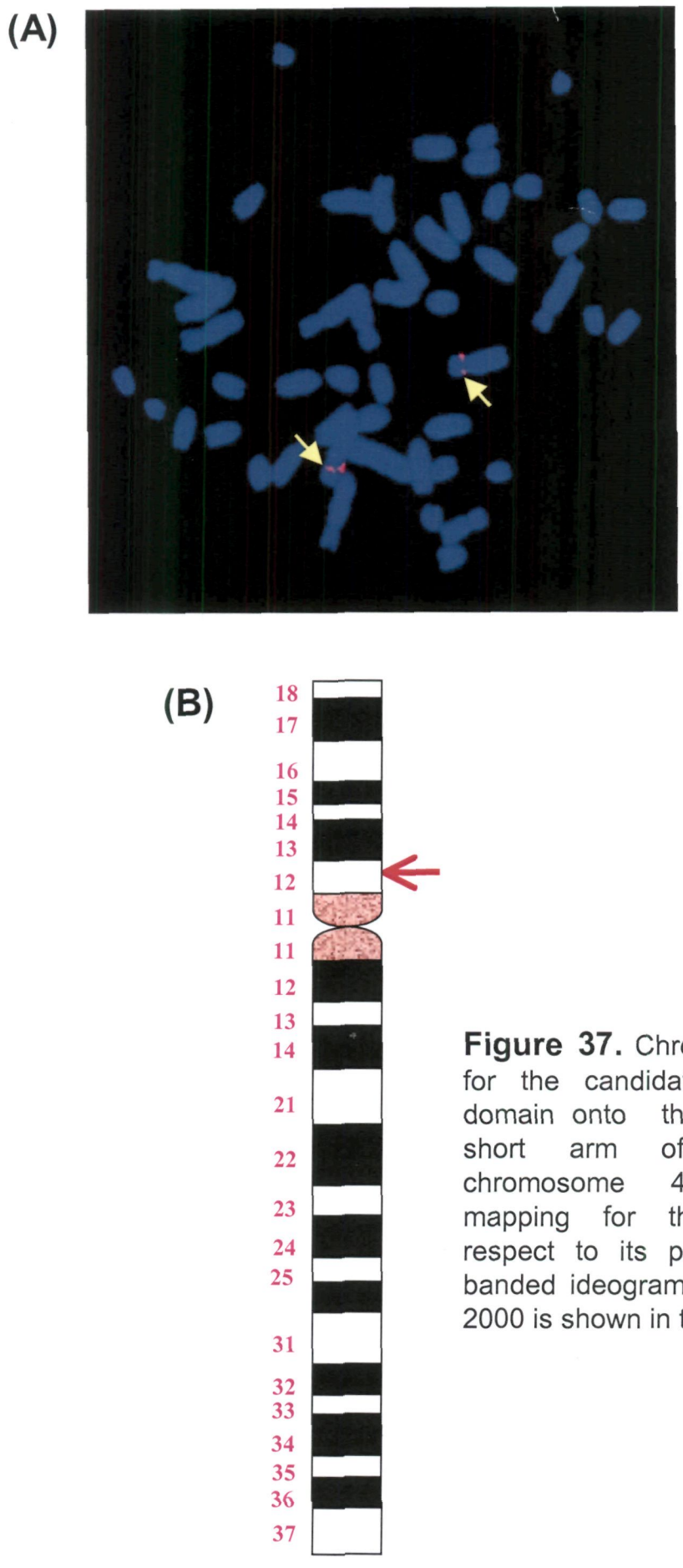


Figure 37. Chromosomal mapping for the candidate Ankyrin repeat domain onto the proximal end of short arm of sub-metacentric chromosome 4 **(A)**. Detailed mapping for these genes with respect to its position on the G-banded ideogram following ISCNDB 2000 is shown in the figure **(B)**.

4.2 Isolation and detailed characterization of the candidate genes

After accessing several known and novel genes tagged with the simple repeats of 33.15, GACA and GATA, we proceeded with the detailed characterization of few candidates for their full length isolation, domain organization, copy number status, *in silico* structural and functional analysis, *in-vitro* protein expression & purification, tissue & age specific transcription/translation and localization of the same onto the metaphase chromosomes.

4.2.1 Proto-oncogene *C-kit* receptor

4.2.1.1 *Isolation of full length CDS of buffalo c-kit receptor*

Following the conserved sequence(s) of *c-kit* across the species, four different primer sets were designed for the buffalo *c-kit* (Table 2), and used for the amplification of different *c-kit* fragments (Figure 38). These *c-kit* fragments were individually cloned, and individual clones were confirmed by restriction analyses (Figure 39) following sequencing of all the fragments. The full length CDS of *c-kit* gene (Accession number: DQ314491) of water buffalo was deduced from the different overlapping fragments (Figure 38A, 40).

The analyses of these different fragments of *c-kit* CDS revealed more than 85% homology at nucleotide level and approximately 95% identity at amino acid level with that of cattle. However, Clone I contained 1497 bp insert with an open reading frame of 1454 bp encompassing immunoglobulin like folds and a small 5'-untranslated region (UTR). Clone II (1449 bp) covered nucleotides from 1525-2973 whereas Clone III represented the intermediate nucleotides 1498-2995 of the complete CDS with the overlapping sequence from clone I and II. Clone IV encompassing nucleotides 825-1504 of the full length *c-kit* CDS was used to confirm the overlapping sequences. This full length CDS encodes a putative protein of about 975 amino acids with a molecular mass of 108.2 kDa (Figure 40 and 41). Comparison of the sequences with human and cattle *c-kit* gene(s)

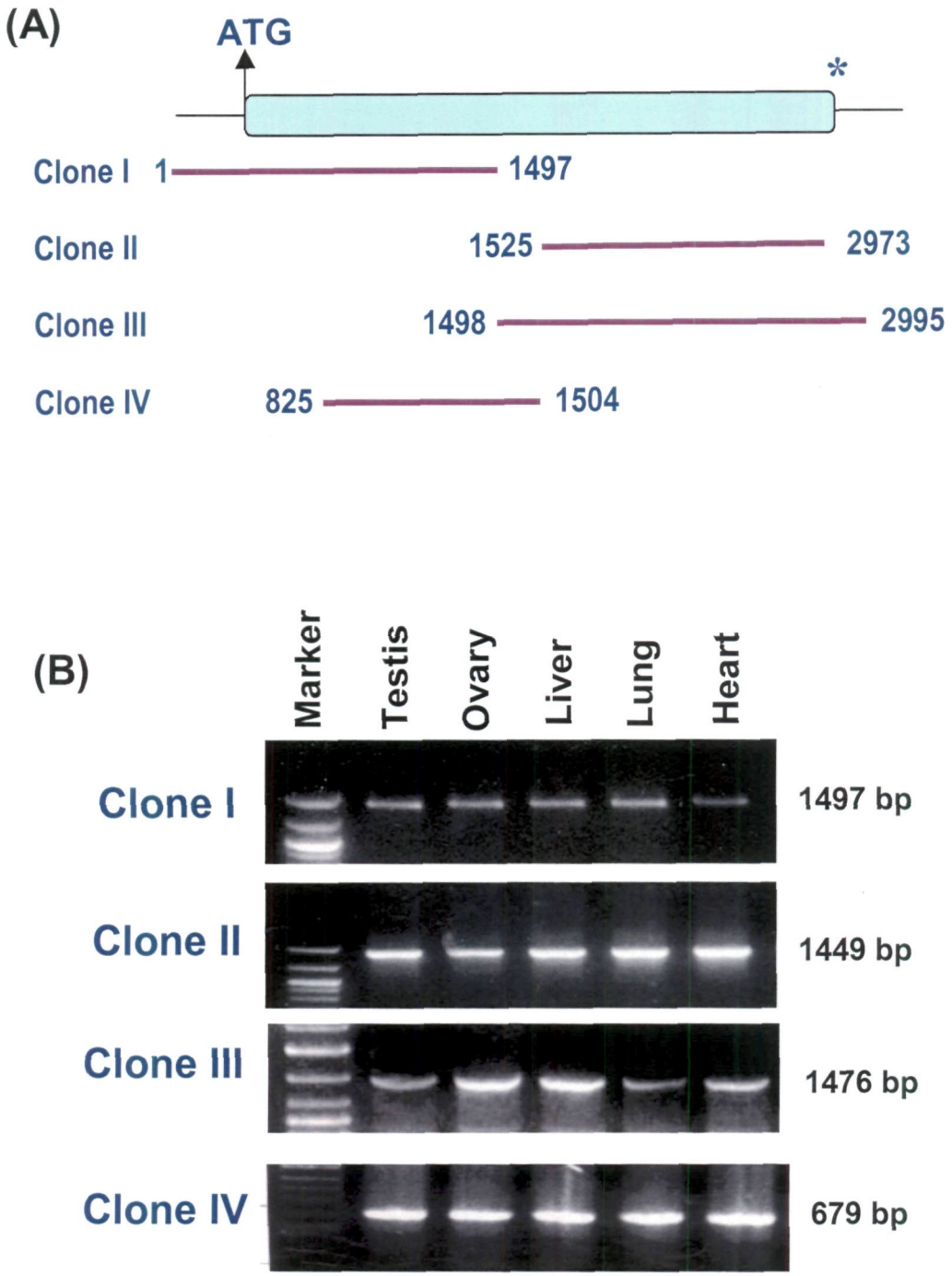


Figure 38. Cloning strategy showing isolation of full length *c-kit* cDNA from buffalo. Different fragments generated by PCR to deduce full length *c-kit* CDS and their nucleotide boundaries are shown in **(A)**. The PCR amplification of these different fragments from different tissues has been shown in **(B)**. ATG are for start codon whereas '*' denotes the stop codon.

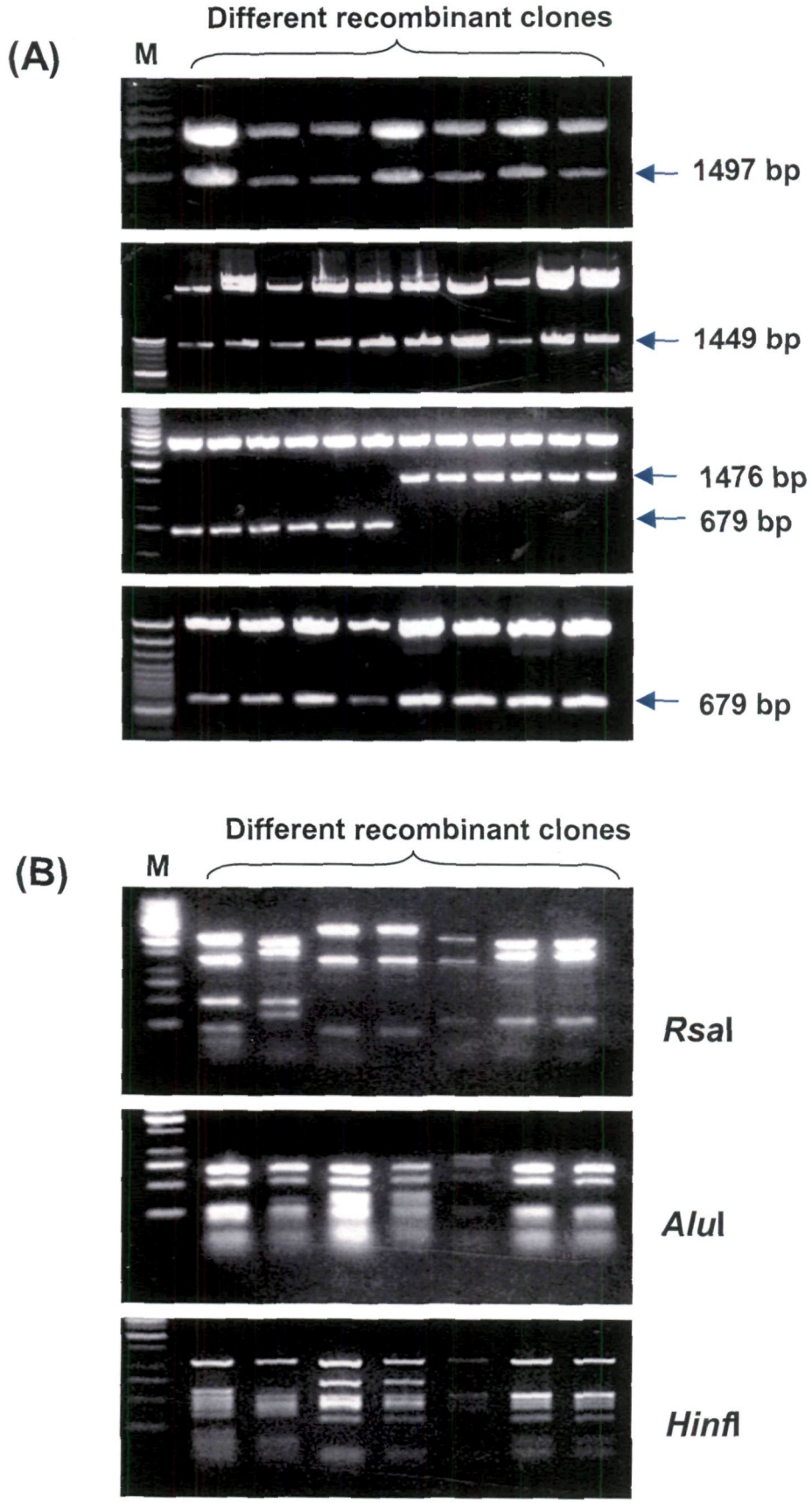


Figure 39. Representative gel pictures for the restriction analyses done to check the insert size from the recombinant clones of *c-kit* using *EcoR*1 **(A)** and for interclonal variation using other enzyme sets **(B).** The sizes of insert released has been shown in the right of each gel picture. No interclonal variation was observed in any of the recombinant clones. The molecular size marker "M" is given in base pairs. .

```
   1  CCGGAACGTGGAACAGAGCTCCGGTCCTAGCGCAGCCACCGCG
  44  ATGAGAGGCGCTCGCGGCGCCTGGGATTTCCTCTTCGTTCTGCTG
      M  R  G  A  R  G  A  W  D  F  L  F  V  L  L
  89  CTCCTGCTCCTCGTCCAGACAGGCTCTTCTCAGCCTTCTGTGAGT
      L  L  L  L  V  Q  T  G  S  S  Q  P  S  V  S
 134  CCAGGGGAACTGTCTCTACCATCTATCCACCCAGCAAAATCAGAG
      P  G  E  L  S  L  P  S  I  H  P  A  K  S  E
 179  TTAATTGTCAGCGTTGGCGACGAGATTAGGCTGTTATGCACTGAT
      L  I  V  S  V  G  D  E  I  R  L  L  C  T  D
 224  CCAGGATTTGTCAAGTGGACTTTTGAGATCCTGGGTCAACTGAGT
      P  G  F  V  K  W  T  F  E  I  L  G  Q  L  S
 269  GAGAAAACAAACCCGGAATGGATCACCGAGAAAGCAGAGGTCACA
      E  K  T  N  P  E  W  I  T  E  K  A  E  V  T
 314  AATACAGGCAATTACACGTGCACCAATAAAGGCGGCTTGAGCAGT
      N  T  G  N  Y  T  C  T  N  K  G  G  L  S  S
 359  TCCATCTATGTGTTTGTTAGAGACCCCGAGAAGCTTTTCCTGATT
      S  I  Y  V  F  V  R  D  P  E  K  L  F  L  I
 404  GACCTTCCCTTGTACGGGAAAGAAGAAAACGACACGCTGGTTCGC
      D  L  P  L  Y  G  K  E  E  N  D  T  L  V  R
 449  TGTCCCCTGACAGACCCCGAGGTGACCAATTACTCTCTCACGGGG
      C  P  L  T  D  P  E  V  T  N  Y  S  L  T  G
 494  TGTGAGGGGAAACCTCTCCCTAAGGATTTGACGTTTGTGGCTGAC
      C  E  G  K  P  L  P  K  D  L  T  F  V  A  D
 539  CCCAAGGCAGGCATCACAATCAGAAATGTGAAGCGTGAGTACCAT
      P  K  A  G  I  T  I  R  N  V  K  R  E  Y  H
 584  CGGCTCTGTCTGCACTGCTCAGCGAATCAGAGGGGCAAGTCCGTG
      R  L  C  L  H  C  S  A  N  Q  R  G  K  S  V
 629  CTGTCGAAGAAATTCACTCTGAAAGTGCGGGCAGCCATCAAAGCT
      L  S  K  K  F  T  L  K  V  R  A  A  I  K  A
 674  GTGCCAGTTGTGTCTGTGTCCAAAACCAGCTATCTTCTCAGGGAA
      V  P  V  V  S  V  S  K  T  S  Y  L  L  R  E
 719  GGAGAGGAATTTGCAGTGACATGCTTGATTAAAGACGTGTCTAGT
      G  E  E  F  A  V  T  C  L  I  K  D  V  S  S
 764  TCCGTGGACTCTATGTGGATAAAGGAAAACAGCCAGCAGACTAAA
      S  V  D  S  M  W  I  K  E  N  S  Q  Q  T  K
 809  GCACAGACGAAGAAGAATAGCTGGCATCAGGGTGACTTCAGTTAT
      A  Q  T  K  K  N  S  W  H  Q  G  D  F  S  Y
 854  CTCCGTCAGGAAAGGTTGACTATCAGCTCAGCAAGAGTGAATGAT
      L  R  Q  E  R  L  T  I  S  S  A  R  V  N  D
 899  TCTGGTGTGTTCATGTGTTACGCCAATAATACTTTTGGATCAGCA
      S  G  V  F  M  C  Y  A  N  N  T  F  G  S  A
 944  AATGTCACAACAACCTTAGAAGTAGTAGATAAAGGATTCATTAAT
      N  V  T  T  T  L  E  V  V  D  K  G  F  I  N
 989  ATCTTCCCTATGATGAACACAACAGTATTTGTAAATGATGGAGAG
      I  F  P  M  M  N  T  T  V  F  V  N  D  G  E
1034  AATGTGGATCTGGTTGTTGAATATGAGGCATATCCCAAACCTGTA
      N  V  D  L  V  V  E  Y  E  A  Y  P  K  P  V
1079  CACCGACAGTGGATATATATGAACAGAACCTCCACTGATAAGTGG
      H  R  Q  W  I  Y  M  N  R  T  S  T  D  K  W
1124  GACGATTATCCCAAGTCTGAAAATGAAAGTAACATCAGATACGTA
      D  D  Y  P  K  S  E  N  E  S  N  I  R  Y  V
1169  AATGAACTTCATCTAACCAGATTAAAAGGGACTGAAGGAGGCACT
      N  E  L  H  L  T  R  L  K  G  T  E  G  G  T
1214  TACACATTTCACGTGTCCAATTCTGATGTCAATTCTTCCGTGACA
      Y  T  F  H  V  S  N  S  D  V  N  S  S  V  T
1259  TTTAACGTTTACGTGAACACAAAACCAGAAATCCTGACGCATGAC
      F  N  V  Y  V  N  T  K  P  E  I  L  T  H  D
1304  AGGCTGGTGAATGGCATGCTACAGTGCGTGGCCGCAGGGTTCCCG
      R  L  V  N  G  M  L  Q  C  V  A  A  G  F  P
1349  GAGCCAACCATCGATTGGTACTTTTGTCCAGGAACCGAGCAGAGG
      E  P  T  I  D  W  Y  F  C  P  G  T  E  Q  R
1394  TGTTCTGTCCCGTTGGGCCAGTGGATGTACAGATCCAAAACTCAT
      C  S  V  P  L  G  Q  W  M  Y  R  S  K  T  H
1439  CTGTCTCACCATTGGAAAACTAGTGGTAAGTGCACCATTGATGAC
      L  S  H  H  W  K  T  S  G  K  C  T  I  D  D
1484  AGCACATTCAAACAAATGGGACGGTGGAGTGCAGGGCTTATAACG
      S  T  F  K  Q  M  G  R  W  S  A  G  L  I  T
1529  ATGGTGGGCAAGAGTTCTGCCTCTTTTAACTTTGTATTTAAAGGT
      M  V  G  K  S  S  A  S  F  N  F  V  F  K  G
```


Figure 40

Contd/-

```
1574 AACAGCAAAGAACAAATCCATGCTCACACCCTGTTCACGCCGTTG
     N  S  K  E  Q  I  H  A  H  T  L  F  T  P  L
1619 CTGATTGGTTTTGTGATCGCAGCTGGTTTAATGTGTATCTTCGTG
     L  I  G  F  V  I  A  A  G  L  M  C  I  F  V
1664 GTGATTCTTACGTACAAATATTTGCAGAAACCCATGTATGAAGTA
     V  I  L  T  Y  K  Y  L  Q  K  P  M  Y  E  V
1709 CAGTGGAAAGTTGTCGAGGAGATAAATGGAAACAATTATGTTTAC
     Q  W  K  V  V  E  E  I  N  G  N  N  Y  V  Y
1754 ATAGACCCAACACAACTTCCTTATGATCACAAATGGGAGTTTCCC
     I  D  P  T  Q  L  P  Y  D  H  K  W  E  F  P
1799 AGGAACAGGCTGAGTTTTGGGAAAACCTTGGGTGCTGGCGCCTTC
     R  N  R  L  S  F  G  K  T  L  G  A  G  A  F
1844 GGGAAAGTTGTTGAGGCCACCGCTTATGGCTTAATTAAATCAGAT
     G  K  V  V  E  A  T  A  Y  G  L  I  K  S  D
1889 GCAGCCATGACTGTTGCTGTCAAGATGCTCAAACCAAGCGCCCAT
     A  A  M  T  V  A  V  K  M  L  K  P  S  A  H
1934 TTAACAGAACGAGGAGCCCTAATGTCTGAACTCAAAGTCTTGAGT
     L  T  E  R  G  A  L  M  S  E  L  K  V  L  S
1979 TACCTCGGTAATCATATGAATATTGTGAATCTTCTGGGAGCGTGC
     Y  L  G  N  H  M  N  I  V  N  L  L  G  A  C
2024 ACCATTGGAGGGCCCACCCTGGTCATTACAGAATATTGTTGCTAT
     T  I  G  G  P  T  L  V  I  T  E  Y  C  C  Y
2069 GGTGACCTTCTGAATTTTTTGAGAAGAAAACGTGATTCATTTATT
     G  D  L  L  N  F  L  R  R  K  R  D  S  F  I
2114 TGCTCAAAGCAGGAAGATCACGCCGAAGTGGCGCTTTATAAGAAC
     C  S  K  Q  E  D  H  A  E  V  A  L  Y  K  N
2159 CTTTCTTCATTCAAAGGAGTCTTCCTGCAATGTTTGTACTATGAG
     L  S  S  F  K  G  V  F  L  Q  C  L  Y  Y  E
2204 TACATGGACATGAAACCTGGAGTTTCTTATGTTGTACCAACCAAG
     Y  M  D  M  K  P  G  V  S  Y  V  V  P  T  K
2249 GCAGACAAGAGGAGATCTGCAAGAATAGGGTTCATACATAGAAAG
     A  D  K  R  R  S  A  R  I  G  F  I  H  R  K
2294 AGACGTGACTCCTGCTATCATGGAAGATGTTTGGCTGGCCCCTGG
     R  R  D  S  C  Y  H  G  R  C  L  A  G  P  W
2339 ACCTGGAGGACTTGCTGCGCTTTTCTTACCAGGTGGCAAAAAGGC
     T  W  R  T  C  C  A  F  L  T  R  W  Q  K  G
2384 ATGGCGTTCCTTGCCTCAAAGAATTGTATTCATAGAGACTTGGCA
     M  A  F  L  A  S  K  N  C  I  H  R  D  L  A
2429 GCCAGAAATATCCTCCTTACTCATGGTCGAATCACAAAGATTTGT
     A  R  N  I  L  L  T  H  G  R  I  T  K  I  C
2474 GATTTTGGTCTCGCCAGAGACATCAAGAATGATTCTAATTATGTG
     D  F  G  L  A  R  D  I  K  N  D  S  N  Y  V
2519 GTCAAAGGAAACGCTCGACTCCCTGTGAAGTGGATGGCACCAGAG
     V  K  G  N  A  R  L  P  V  K  W  M  A  P  E
2564 AGTATTTTCAACTGTGCTACCAGTGCTCTCCTTGCTGATTGTCAT
     S  I  F  N  C  A  T  S  A  L  L  A  D  C  H
2609 GCTGACTTGCAAACTGTTTGTGCCTCAGGAAGCAGCCCCTACCCT
     A  D  L  Q  T  V  C  A  S  G  S  S  P  Y  P
2654 GGAATGCCAGTCGATTCTAAGTTCTACAAGATGATCAAGGAAGGT
     G  M  P  V  D  S  K  F  Y  K  M  I  K  E  G
2699 TTCCGAATGCTCAGCCCCGAGCATGCACCTGCGGAAATGTATGAC
     F  R  M  L  S  P  E  H  A  P  A  E  M  Y  D
2744 ATCATGAAGACCTGCTGGGATGCTGATCCCTTGAAAAGGCCAACA
     I  M  K  T  C  W  D  A  D  P  L  K  R  P  T
2789 TTTAAGCAGATTGTGCAGCTGATTGAGAAGCAGATCTCAGAGAGC
     F  K  Q  I  V  Q  L  I  E  K  Q  I  S  E  S
2834 ACCAATCATATTTATTCCAACTTAGCAAACTGCAGTCCCCACCGG
     T  N  H  I  Y  S  N  L  A  N  C  S  P  H  R
2879 GAGAACCCCACCGTGGACCATTCTGTGCGCATCAACTCTGTGGGC
     E  N  P  T  V  D  H  S  V  R  I  N  S  V  G
2924 AGCAGCGCCTCCTCCACGCAGCCTCTGCTTGTCCACGAAGATGTC
     S  S  A  S  S  T  Q  P  L  L  V  H  E  D  V
2969 TGATC 2973
     *
```

Figure 40. Complete *c-kit* cDNA sequence along with the deduced amino acids and exons are given. The exons are shown in alternate colors and the start/stop codons are overshadowed.

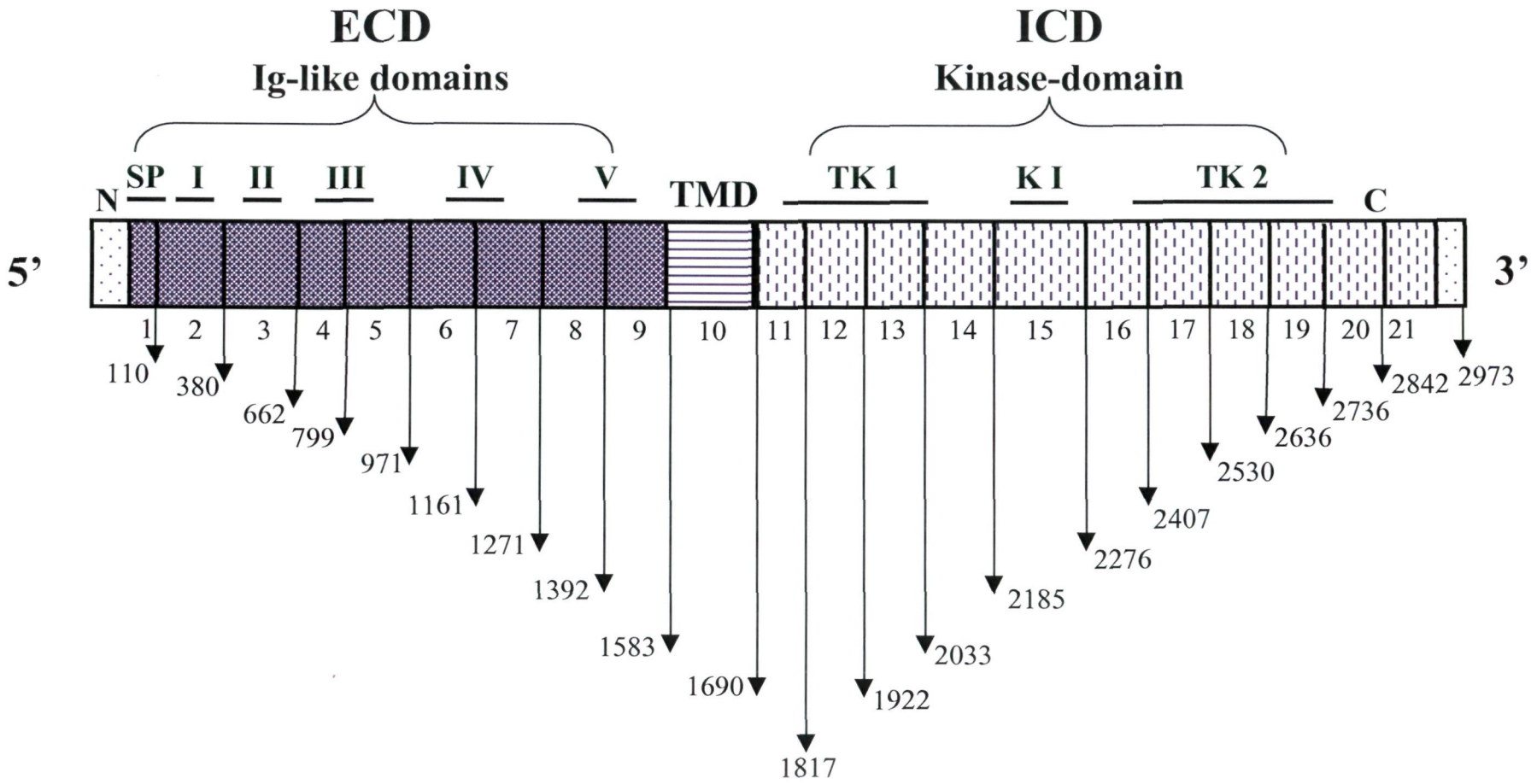


Figure 41. Diagrammatic illustration of *c-kit* gene representing the domain organization and nucleotide boundary of each exon. The SP stands for signal peptide; ECD, Extra cellular domain; ICD, Intracellular domain; TMD, Trans-membrane domain; TK1/TK2 represents tyrosine kinase domain 1 & 2; and KI for Kinase insert.

facilitated delineation of the exon-intron boundaries. Buffalo *c-kit* gene has 21 exons with exon/intron boundary similar to those reported in the cattle, goat and human. The exon boundaries and their encoding domains have been demonstrated in the figure 41).

4.2.1.2 *Domain organization of buffalo c-kit receptor*

The homology search of the c-kit protein showed the presence of all the domains characteristics of the receptor tyrosine kinase (RTK) family (Figure 41 and 42). Of the amplified 2973 bp of *c-kit* gene, 44-2969 nucleotides encode 975 amino acids. Exons 1-9 encode extracellular domain containing immunoglobulin folds with 5'UTR and signal peptide, exon 10 codes for the transmembrane region while 11-21 represented the cytoplasmic domain consisting of tyrosine kinase and 3'UTR regions (Figure 42). Of the full length c-kit peptide, first 27 amino acids conformed to the signal peptide consensus whereas "aa" 220 to 308 represented Immunoglobulin like folds (Figure 42 and 43) and C-terminal region from residues 581-925 showed homology with catalytic tyrosine kinase domain and both domains have been separated by a small transmembrane domain. The tyrosine kinase domain harbors a conserved catalytic core common to both serine/threonine and tyrosine protein kinases involved in signal transduction (Figure 42 and 43).

4.2.1.3 *Tissue specific sequence variation resulting in truncated peptides*

Sequence alignment of *c-kit* gene across different somatic and gonadal tissues of buffalo showed several point nucleotide changes and INDELS specific to a particular tissue (Figure 44). These changes included a major deletion of 12 bp in lung and several single nucleotide deletions/substitutions in other tissues (Figure 44) resulting in the truncated peptide(s) of different lengths (Figure 45). Significantly, the full length c-kit protein of 975 residues was translated only in testis whereas remaining tissues showed either absence of the intracellular or transmembrane domains or both. This gave rise to a novel truncated c-kit containing only extracellular domain. The peptide length varied among

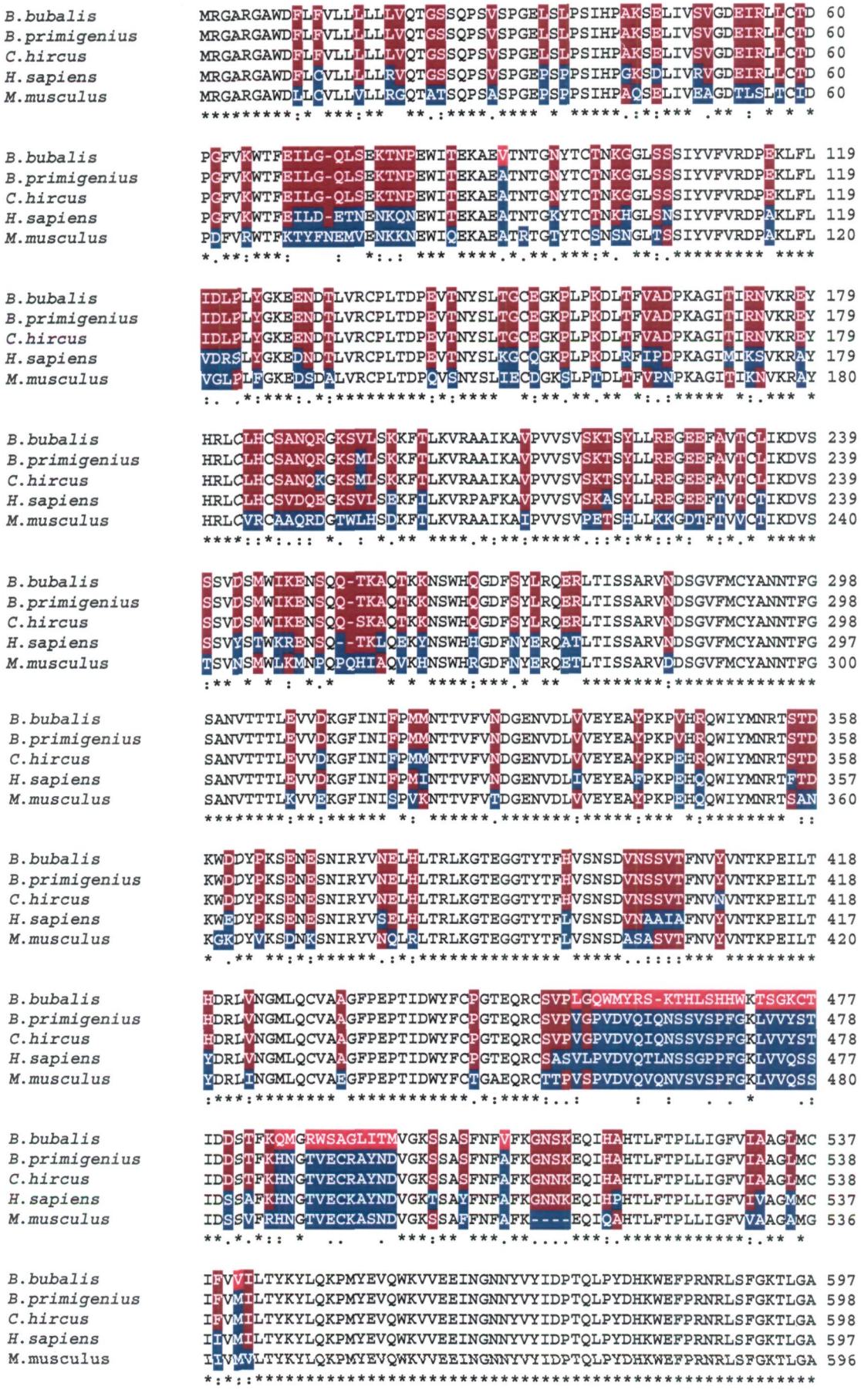


Figure 42

Contd/-

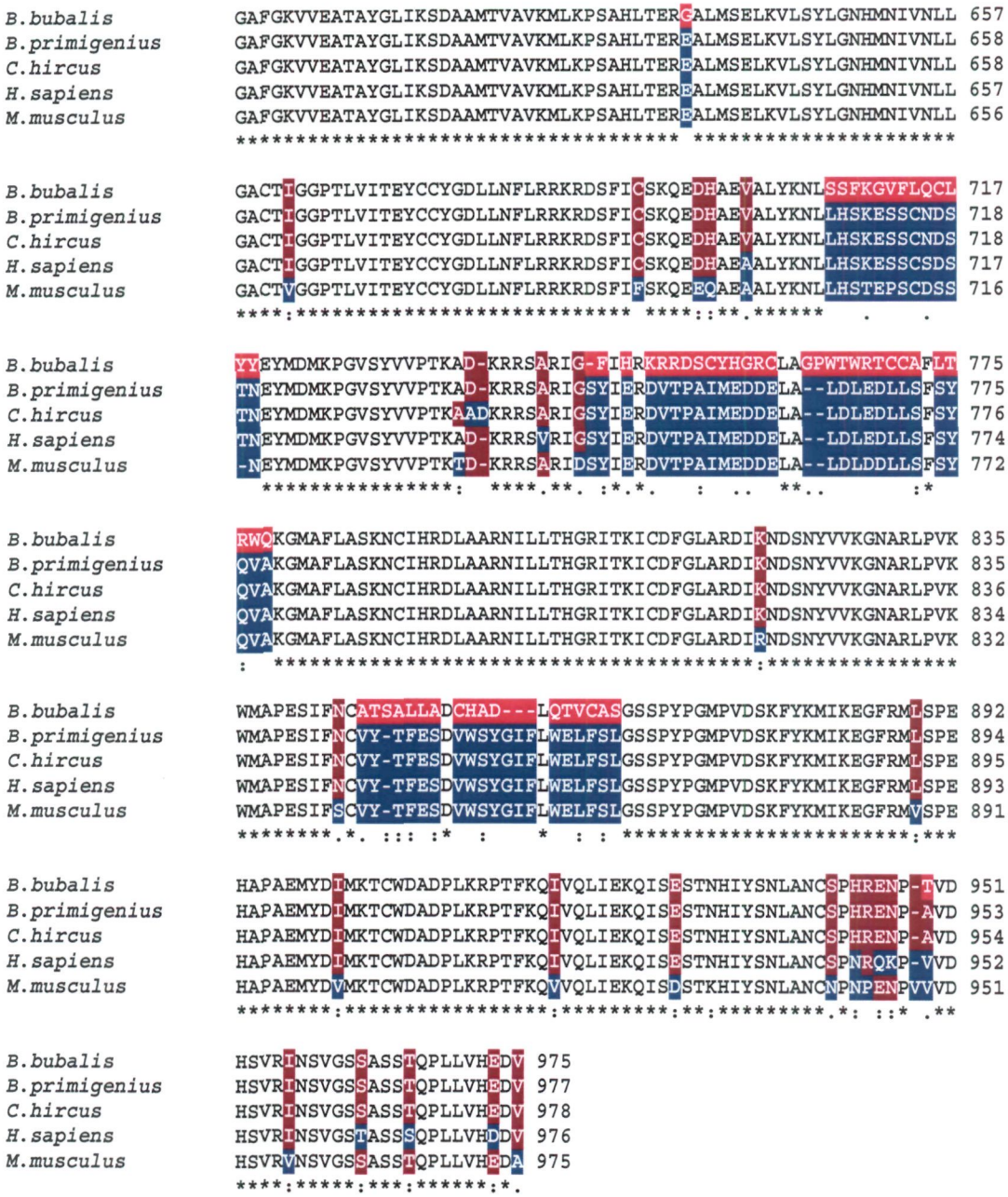
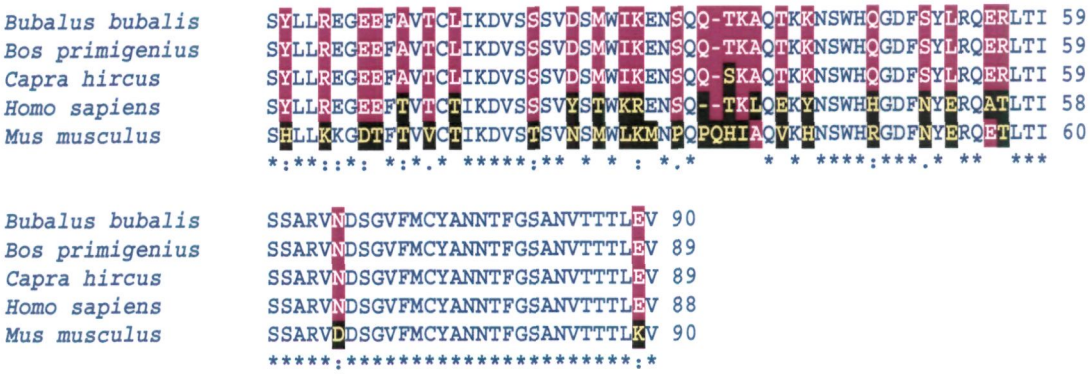


Figure 42. Amino acid sequence alignment of buffalo c-kit with that of different species. Mutational hotspots exclusive to buffalo c-kit have been highlighted in red. The variations in buffalo shared by the cattle and goat has been shadowed dark red while the ones shared by the human are in blue background.

(A) Immunoglobulin like domain



(B) Tyrosine kinase domain

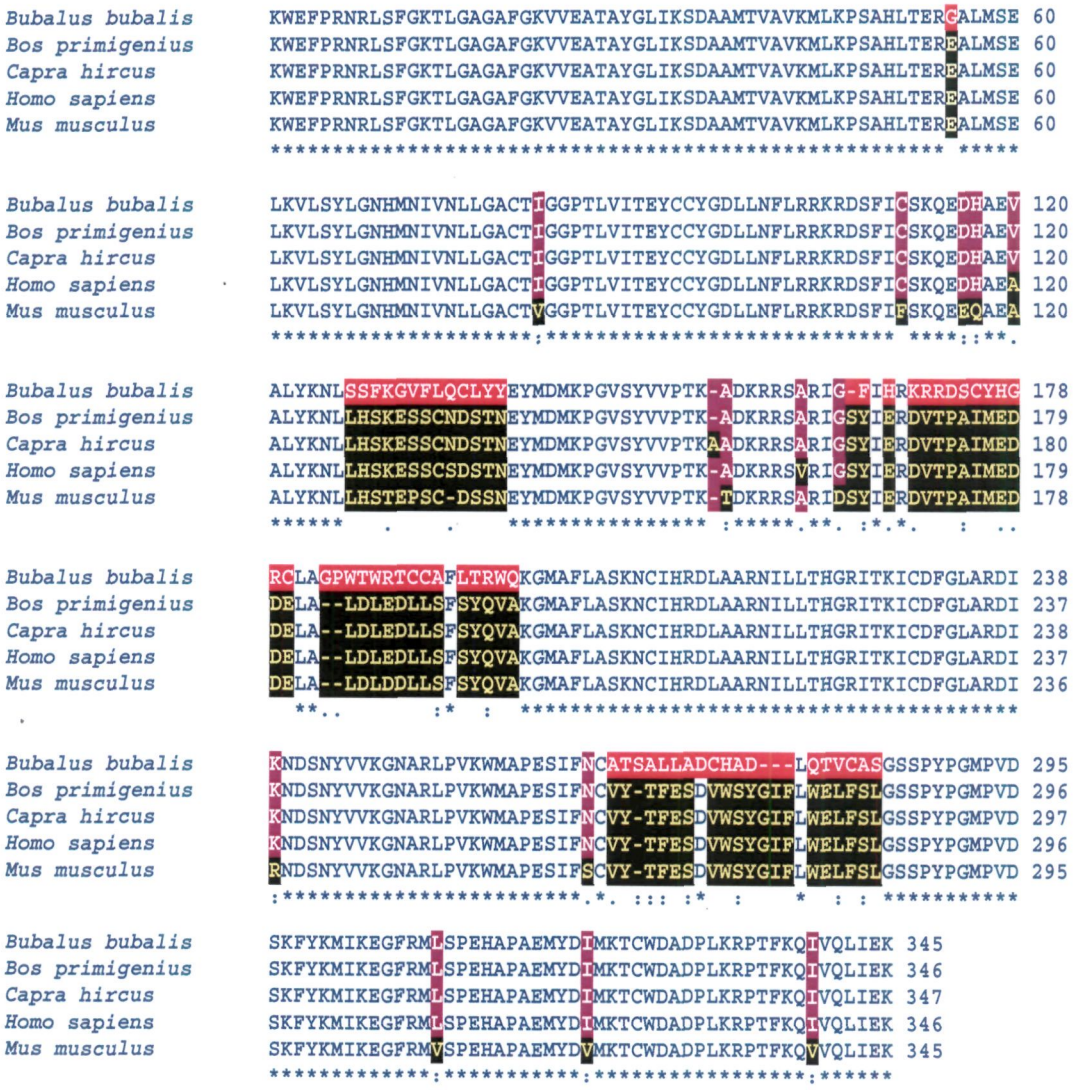


Figure 43. Amino acid alignment of immunoglobulin **(A)** and tyrosine kinase **(B)** domains of c-kit from different species. Mutational hotspots in the buffalo c-kit at three places (highlighted red) were detected compared to that in other species.

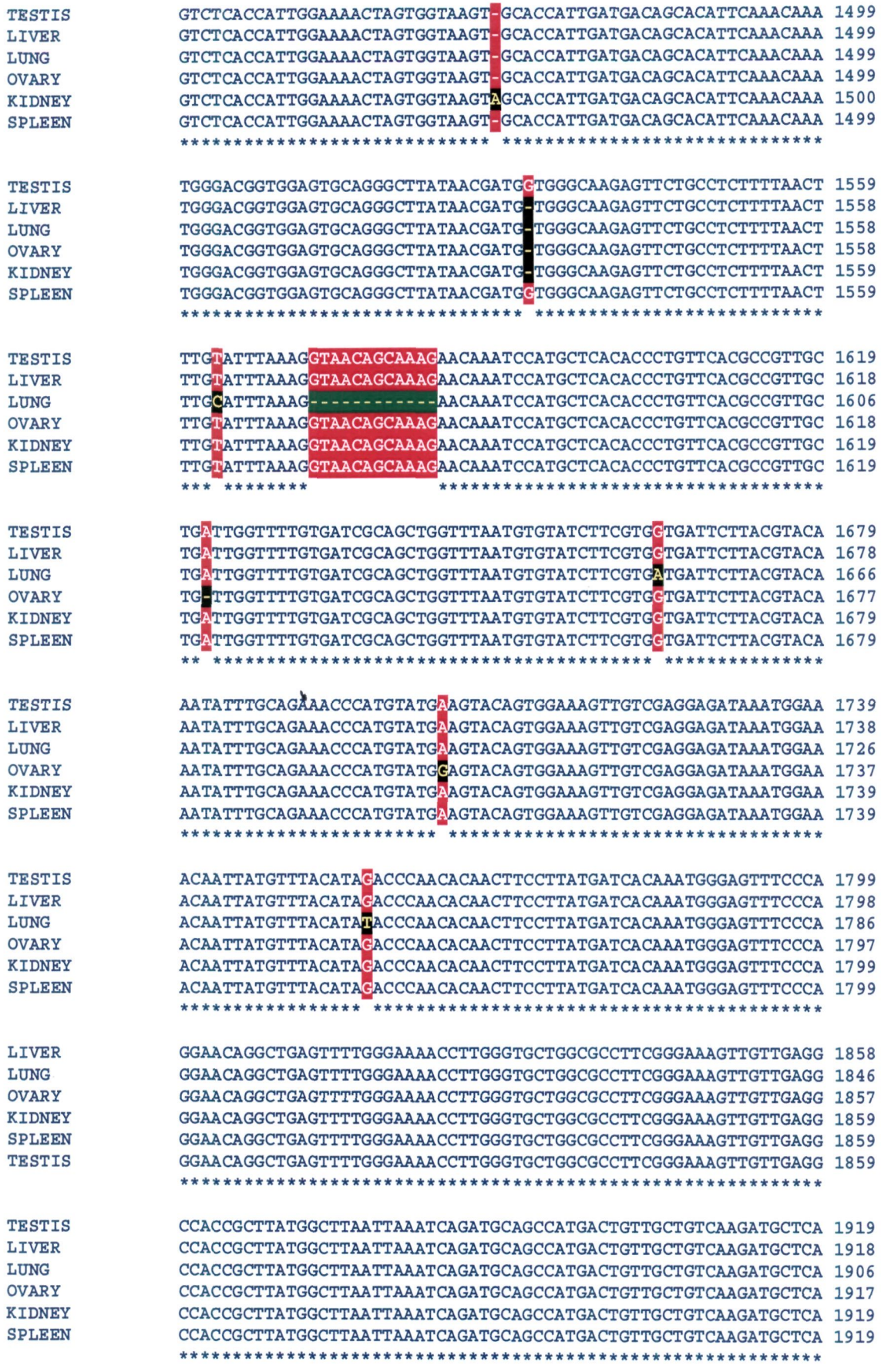


Figure 44

Contd/-

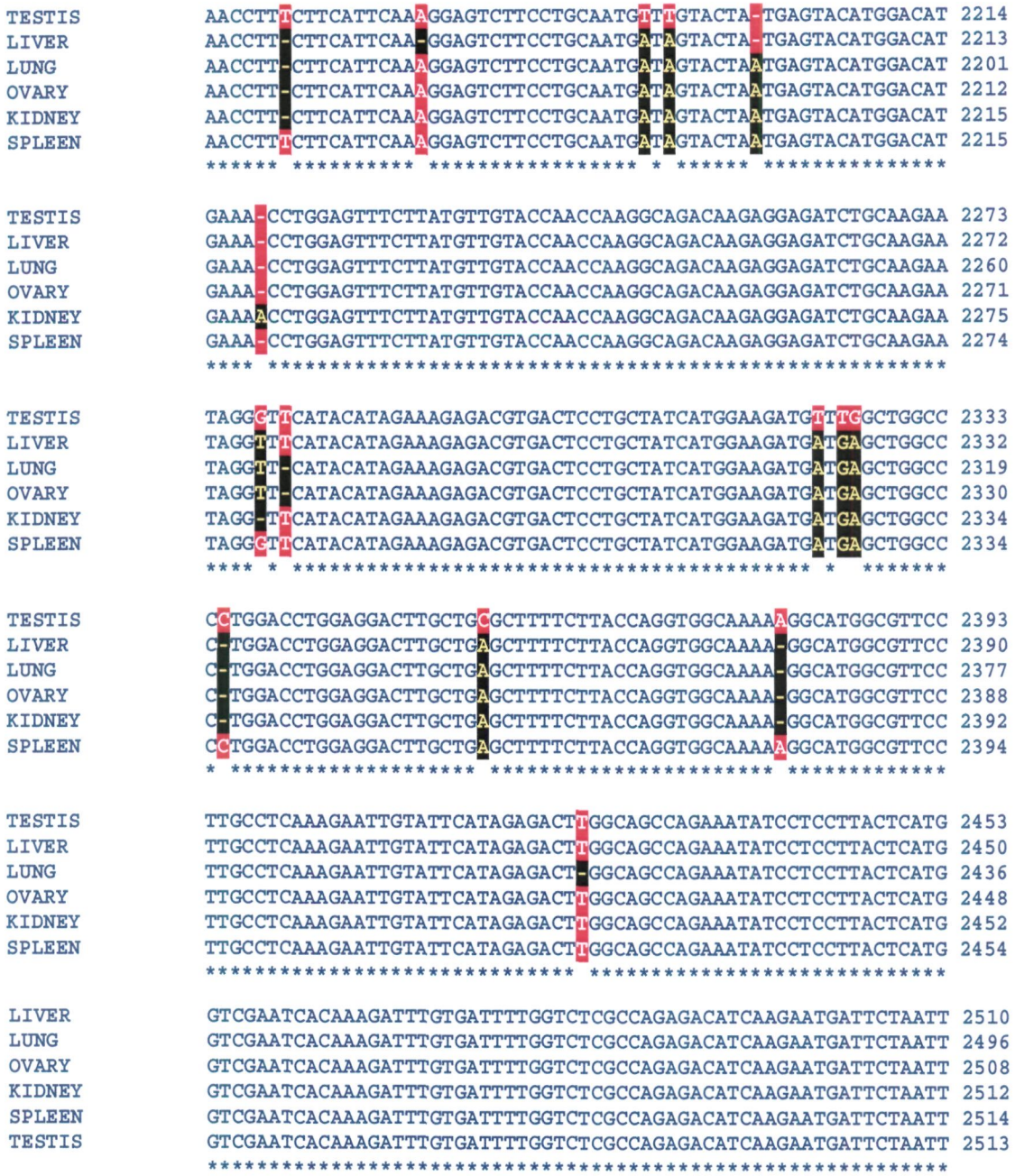


Figure 44. Multiple nucleotide sequence alignment showing the c-kit region prone for the nucleotide changes amongst different tissues of adult buffalo. Please note the 12 bp deletion in the lung and several other point nucleotide variations compared to that in other tissues, which are highlighted. This was further confirmed by using samples from five additional animals.

different tissues showing 475 residues in kidney, 525 in liver, 521 in lung, 577 in ovary and 674 in spleen (for details, see Figure 45).

4.2.1.4 *Alternate splicing of buffalo c-kit in different tissues*

The primers designed for amplifying the intracellular and transmembrane domains gave rise to many tissue specific alternatively spliced products (Figure 46). Of the four primer pairs used, CK174 and CK175, specific to intracellular domain showed 618 bp amplicon along with the expected 1480 bp one in testis (Figure 46A). In the heart and ovary, the same primer amplified 618 bp and 500 bp fragments, respectively, whereas in kidney and lung, only 500 bp fragment was detected in place of expected 1500 bp (not shown). These were construed to be the alternate transcripts after confirming their presence with the gradients of primer Tm and concentrations of PCR reagents. The presence of faint amplicons was confirmed by hybridizing the dried gel with [P^{32}] α-dCTP labeled *c-kit* probe (Panels b of the figure 46). Other primer pair, CK150 & 151 specific to transmembrane domain and the part of ECD and ICD showed amplification of 480 bp along with the expected 679 bp one in heart, testis and ovary (Figure 46B). This was also confirmed by Southern blotting in the form of discernible signal intensity (panel b of the figure 46B). In lung, the same primer set amplified 250, 500 and 600 bp fragments along with the 683 bp amplicon (not shown). Similarly, the primer set CK148/149 amplified two extra bands of 620 and 350 bp in lung and liver, respectively and a 620 bp one in addition to the expected 1437 bp one in both the heart and spleen (Figure 46C). Primer set CK176/177 specific to extracellular domain did not show alternative splicing.

4.2.1.5 *Uniqueness in the tyrosine kinase domain of buffalo c-kit receptor*

The comparison of the amino acid sequences of the buffalo *c-kit* across the species showed 98% homology with cattle and goat and 88% and 81% with human and mouse, respectively. The *c-kit* CDS from 14 other species and their accession numbers, nucleotide length, amino acid

```
TESTIS      MRGARGAWDFLFVLLLLLLVQTGSSQPSVSPGELSLPSIHPAKSELIVSVGDEIRLLCTD 60
LIVER       MRGARGAWDFLFVLLLLLLVQTGSSQPSVSPGELSLPSIHPAKSELIVSVGDEIRLLCTD 60
OVARY       MRGARGAWDFLFVLLLLLLVQTGSSQPSVSPGELSLPSIHPAKSELIVSVGDEIRLLCTD 60
LUNG        MRGARGAWDFLFVLLLLLLVQTGSSQPSVSPGELSLPSIHPAKSELIVSVGDEIRLLCTD 60
KIDNEY      MRGARGAWDFLFVLLLLLLVQTGSSQPSVSPGELSLPSIHPAKSELIVSVGDEIRLLCTD 60
SPLEEN      MRGARGAWDFLFVLLLLLLVQTGSSQPSVSPGELSLPSIHPAKSELIVSVGDEIRLLCTD 60
            ************************************************************

TESTIS      PGFVKWTFEILGQLSEKTNPEWITEKAEVTNTGNYTCTNKGGLSSSIYVFVRDPEKLFLI 120
LIVER       PGFVKWTFEILGQLSEKTNPEWITEKAEVTNTGNYTCTNKGGLSSSIYVFVRDPEKLFLI 120
OVARY       PGFVKWTFEILGQLSEKTNPEWITEKAEVTNTGNYTCTNKGGLSSSIYVFVRDPEKLFLI 120
LUNG        PGFVKWTFEILGQLSEKTNPEWITEKAEVTNTGNYTCTNKGGLSSSIYVFVRDPEKLFLI 120
KIDNEY      PGFVKWTFEILGQLSEKTNPEWITEKAEVTNTGNYTCTNKGGLSSSIYVFVRDPEKLFLI 120
SPLEEN      PGFVKWTFEILGQLSEKTNPEWITEKAEVTNTGNYTCTNKGGLSSSIYVFVRDPEKLFLI 120
            ************************************************************

TESTIS      DLPLYGKEENDTLVRCPLTDPEVTNYSLTGCEGKPLPKDLTFVADPKAGITIRNVKREYH 180
LIVER       DLPLYGKEENDTLVRCPLTDPEVTNYSLTGCEGKPLPKDLTFVADPKAGITIRNVKREYH 180
OVARY       DLPLYGKEENDTLVRCPLTDPEVTNYSLTGCEGKPLPKDLTFVADPKAGITIRNVKREYH 180
LUNG        DLPLYGKEENDTLVRCPLTDPEVTNYSLTGCEGKPLPKDLTFVADPKAGITIRNVKREYH 180
KIDNEY      DLPLYGKEENDTLVRCPLTDPEVTNYSLTGCEGKPLPKDLTFVADPKAGITIRNVKREYH 180
SPLEEN      DLPLYGKEENDTLVRCPLTDPEVTNYSLTGCEGKPLPKDLTFVADPKAGITIRNVKREYH 180
            ************************************************************

TESTIS      RLCLHCSANQRGKSVLSKKFTLKVRAAIKAVPVVSVSKTSYLLREGEEFAVTCLIKDVSS 240
LIVER       RLCLHCSANQRGKSVLSKKFTLKVRAAIKAVPVVSVSKTSYLLREGEEFAVTCLIKDVSS 240
OVARY       RLCLHCSANQRGKSVLSKKFTLKVRAAIKAVPVVSVSKTSYLLREGEEFAVTCLIKDVSS 240
LUNG        RLCLHCSANQRGKSVLSKKFTLKVRAAIKAVPVVSVSKTSYLLREGEEFAVTCLIKDVSS 240
KIDNEY      RLCLHCSANQRGKSVLSKKFTLKVRAAIKAVPVVSVSKTSYLLREGEEFAVTCLIKDVSS 240
SPLEEN      RLCLHCSANQRGKSVLSKKFTLKVRAAIKAVPVVSVSKTSYLLREGEEFAVTCLIKDVSS 240
            ************************************************************

TESTIS      SVDSMWIKENSQQTKAQTKKNSWHQGDFSYLRQERLTISSARVNDSGVFMCYANNTFGSA 300
LIVER       SVDSMWIKENSQQTKAQTKKNSWHQGDFSYLRQERLTISSARVNDSGVFMCYANNTFGSA 300
OVARY       SVDSMWIKENSQQTKAQTKKNSWHQGDFSYLRQERLTISSARVNDSGVFMCYANNTFGSA 300
LUNG        SVDSMWIKENSQQTKAQTKKNSWHQGDFSYLRQERLTISSARVNDSGVFMCYANNTFGSA 300
KIDNEY      SVDSMWIKENSQQTKAQTKKNSWHQGDFSYLRQERLTISSARVNDSGVFMCYANNTFGSA 300
SPLEEN      SVDSMWIKENSQQTKAQTKKNSWHQGDFSYLRQERLTISSARVNDSGVFMCYANNTFGSA 300
            ************************************************************

TESTIS      NVTTTLEVVDKGFINIFPMMNTTVFVNDGENVDLVVEYEAYPKPVHRQWIYMNRTSTDKW 360
LIVER       NVTTTLEVVDKGFINIFPMMNTTVFVNDGENVDLVVEYEAYPKPVHRQWIYMNRTSTDKW 360
OVARY       NVTTTLEVVDKGFINIFPMMNTTVFVNDGENVDLVVEYEAYPKPVHRQWIYMNRTSTDKW 360
LUNG        NVTTTLEVVDKGFINIFPMMNTTVFVNDGENVDLVVEYEAYPKPVHRQWIYMNRTSTDKW 360
KIDNEY      NVTTTLEVVDKGFINIFPMMNTTVFVNDGENVDLVVEYEAYPKPVHRQWIYMNRTSTDKW 360
SPLEEN      NVTTTLEVVDKGFINIFPMMNTTVFVNDGENVDLVVEYEAYPKPVHRQWIYMNRTSTDKW 360
            ************************************************************

TESTIS      DDYPKSENESNIRYVNELHLTRLKGTEGGTYTFHVSNSDVNSSVTFNVYVNTKPEILTHD 420
LIVER       DDYPKSENESNIRYVNELHLTRLKGTEGGTYTFHVSNSDVNSSVTFNVYVNTKPEILTHD 420
OVARY       DDYPKSENESNIRYVNELHLTRLKGTEGGTYTFHVSNSDVNSSVTFNVYVNTKPEILTHD 420
LUNG        DDYPKSENESNIRYVNELHLTRLKGTEGGTYTFHVSNSDVNSSVTFNVYVNTKPEILTHD 420
KIDNEY      DDYPKSENESNIRYVNELHLTRLKGTEGGTYTFHVSNSDVNSSVTFNVYVNTKPEILTHD 420
SPLEEN      DDYPKSENESNIRYVNELHLTRLKGTEGGTYTFHVSNSDVNSSVTFNVYVNTKPEILTHD 420
            ************************************************************

TESTIS      RLVNGMLQCVAAGFPEPTIDWYFCPGTEQRCSVPLGQWMYRSKTHLSHHWKTSGKCTIDD 480
LIVER       RLVNGMLQCVAAGFPEPTIDWYFCPGTEQRCSVPLGQWMYRSKTHLSHHWKTSGKCTIDD 480
OVARY       RLVNGMLQCVAAGFPEPTIDWYFCPGTEQRCSVPLGQWMYRSKTHLSHHWKTSGKCTIDD 480
LUNG        RLVNGMLQCVAAGFPEPTIDWYFCPGTEQRCSVPLGQWMYRSKTHLSHHWKTSGKCTIDD 480
KIDNEY      RLVNGMLQCVAAGFPEPTIDWYFCPGTEQRCSVPLGQWMYRSKTHLSHHWKTSGK----- 475
SPLEEN      RLVNGMLQCVAAGFPEPTIDWYFCPGTEQRCSVPLGQWMYRSKTHLSHHWKTSGKCTIDD 480
            ********************************************************  :

TESTIS      STFKQMGRWSAGLITMVGKSSASFNFVFKGNSKEQIHAHTLFTPLLIGFVIAAGLMCIFV 540
LIVER       STFKQMGRWSAGLITMWARVLPLLTLYLKVTAK----NKSMLTP-------CSRRC---- 525
OVARY       STFKQMGRWSAGLITMWARVLPLLTLYLKVTAK----NKSMLTP-------CSRRCCWFC 529
LUNG        STFKQMGRWSAGLITMWARVLPLLTLHLK--------NKSMLTP-------CSRRC---- 521
KIDNEY      ------------------------------------------------------------
SPLEEN      STFKQMGRWSAGLITMVGKSSASFNFVFKGNSKEQIHAHTLFTPLLIGFVIAAGLMCIFV 540
```

Figure 45

Contd/-

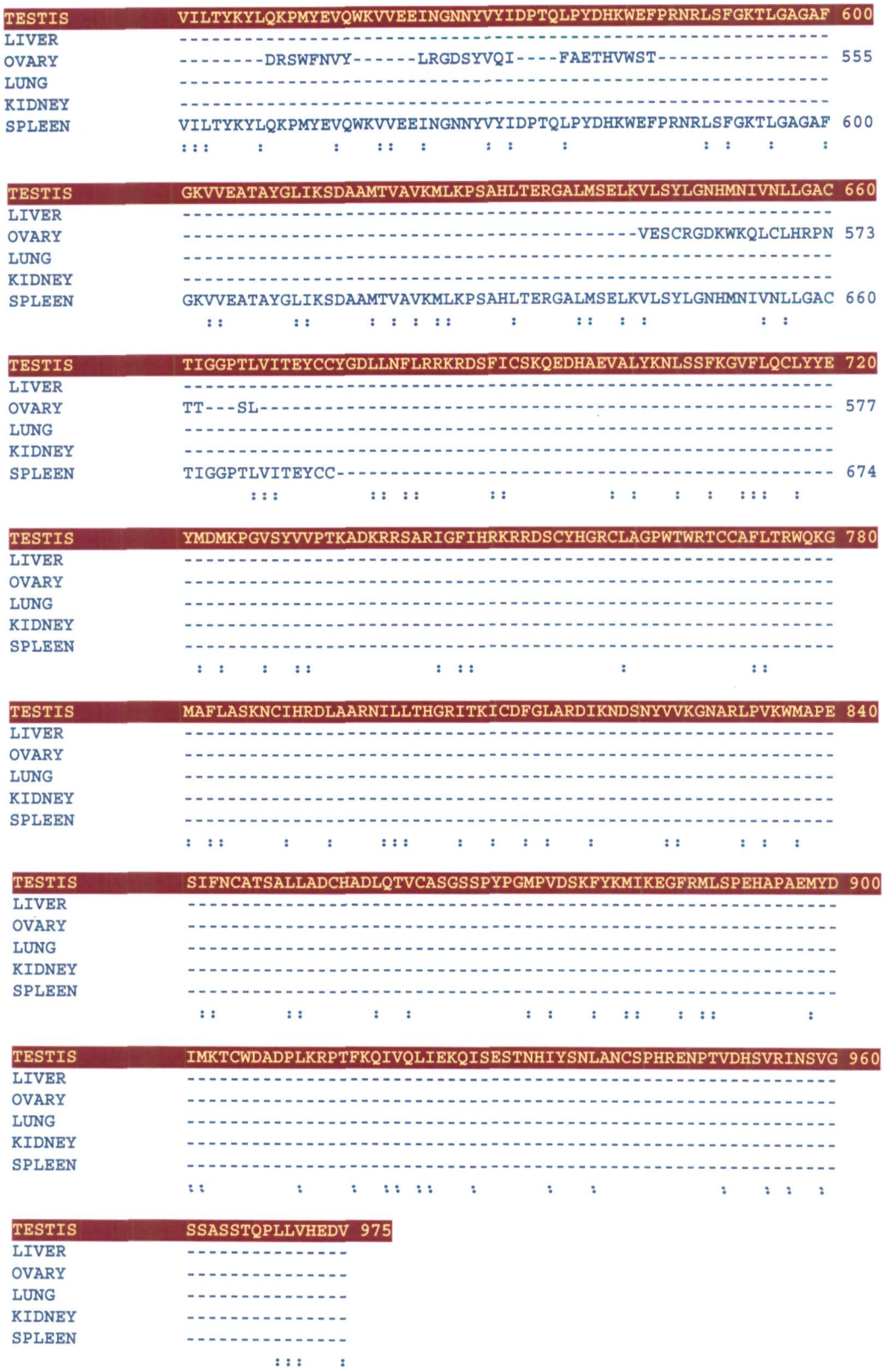


Figure 45. Multiple sequence alignment showing truncated peptides of c-kit from different tissues of adult buffalo except testis. This is due to the lots pf variations in the nucleotides originating from different tissues leading to the stop codon insertion.

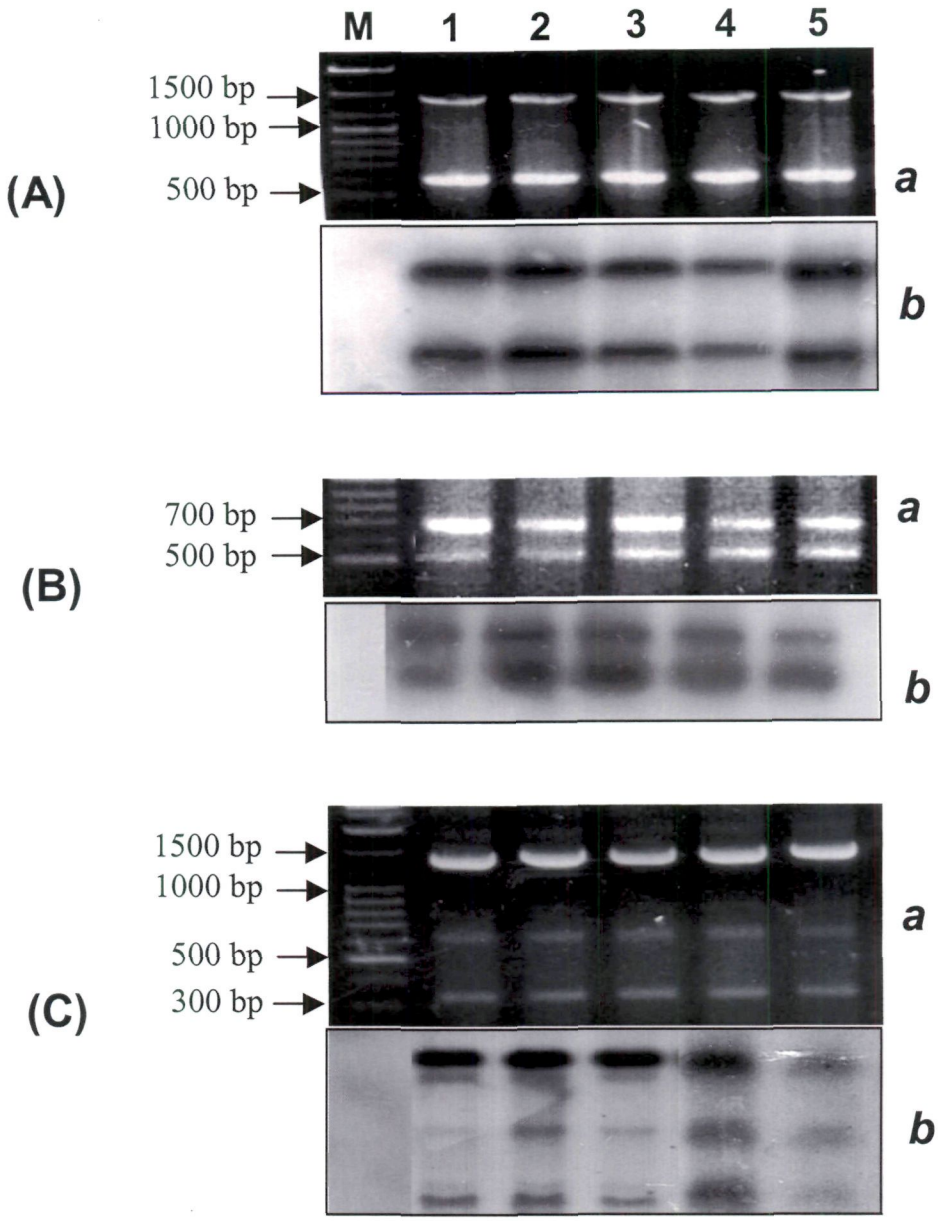


Figure 46. Alternatively spliced mRNA transcripts of c-kit detected with three different sets of primers. Set CK174 and CK175 amplified 618 bp fragment in addition to the expected 1480 bp one in testis **(A),** CK150 & 151 generated 480 bp in addition to the expected 679 bp one in testis **(B)**. Transcripts of similar sizes were also obtained from heart and ovary (not shown). CK148 & 49 gave rise to 620 and 350 bp fragments in addition to the expected 1437 bp one in lung **(C)** and same result was obtained in liver also (not shown). 'M' represents the molecular marker in base pairs, lanes 1-5 represents 5 different annealing temperatures (61-66^{0}C), and sub-panels 'a' and 'b' represents agarose gel and corresponding autoradiogram, respectively, in panels **A-C**.

residues, number of exons, chromosomal position and homology status with buffalo *c-kit* are given in the table 15. Although the IgG like domain of buffalo *c-kit* gene did not show such mutational hotspots, most of the alterations were shared in buffalo, goat and cattle in comparison to human and mice (Figure 43A). Interestingly buffalo *c-kit* gene also showed frequent mutational hotspots in the tyrosine kinase domain making it unique compared to that in goat, cattle, mouse and human *c-kit* gene (Figure 43B). Major amino acid changes included either from a charged polar residue to uncharged polar or from charged/uncharged polar to non-polar residues and *vice-versa* (Figure 43B).

The predicted secondary structure(s) of the buffalo c-kit also showed minor alterations (Figure 47). These changes included a coil of 12 residues in place of helix at one place in the tyrosine kinase domain and random changes throughout the peptide length similar to those detected in goat, cattle or human (Figure 47) without any major impact on the tertiary structure of the protein(s) (Figure 48). The tertiary structures were predicted using (http://www.sbg.bio.ic.ac.uk/phyre).

4.2.1.6 *Evolutionary relationship of buffalo c-kit gene with other species*

The buffalo *c-kit* showed cross-hybridization with genomic DNA from 13 other species with almost equal signal intensity (Figure 49A) confirming its faithful evolutionary conservation in different mammalian species. The *c-kit* sequence(s) were studied *in-silico* among different species for its gene structure, size, translation length and domain organization (For details see Table 15) Phylogenetic trees based on multiple alignment of the *c-kit* sequences from different species (Table 15) showed that the buffalo *c-kit* was closer to that of *Bos taurus, Bos primigenius and Capra hircus* compared to that in other species (Figure 49B). On the other hand, *Gallus gallus* was found to be the most distant species. Only those species were considered for the phylogenetic delineation studied which showed >85% homology with the buffalo *c-kit*.

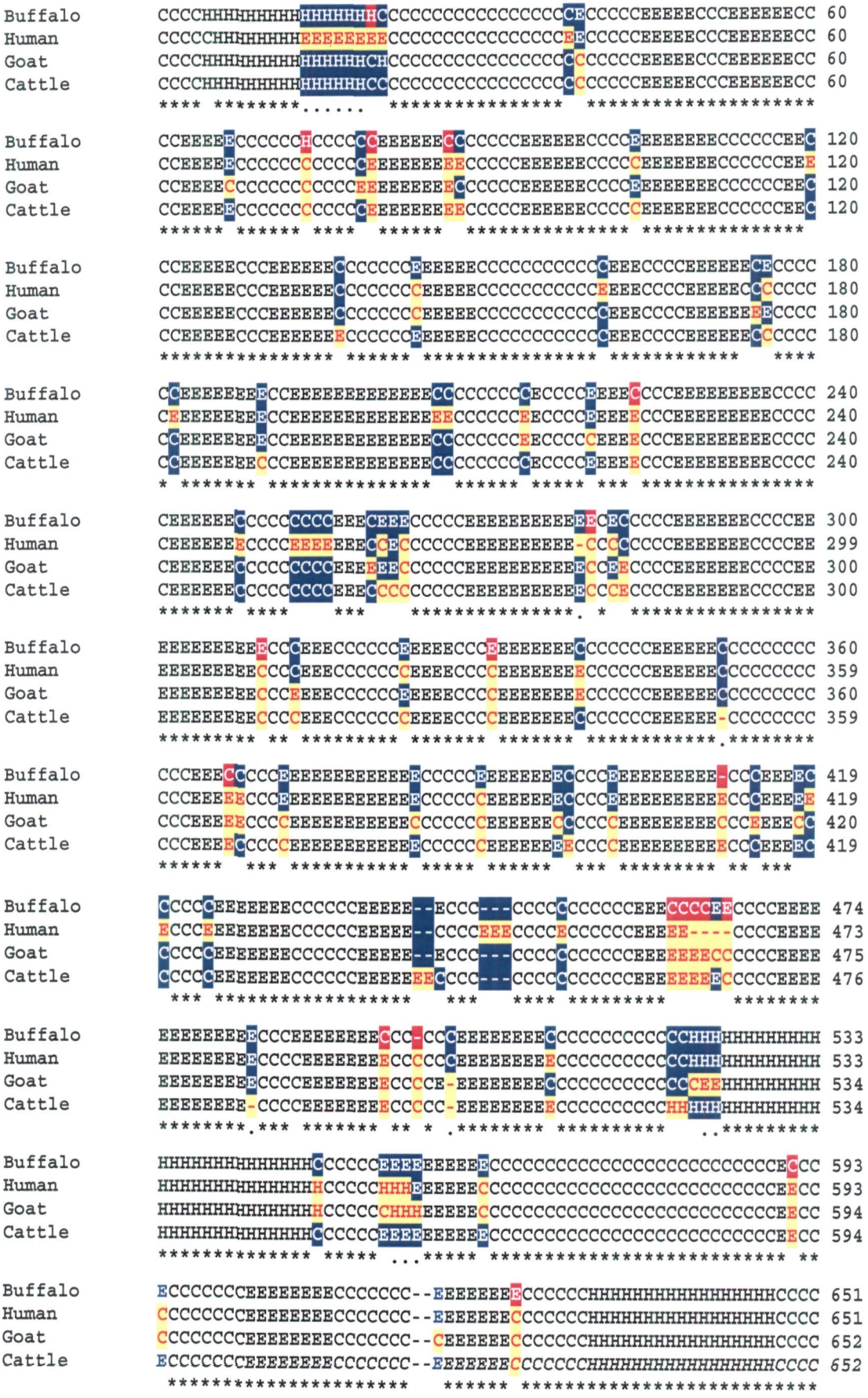


Figure 47

Contd/-

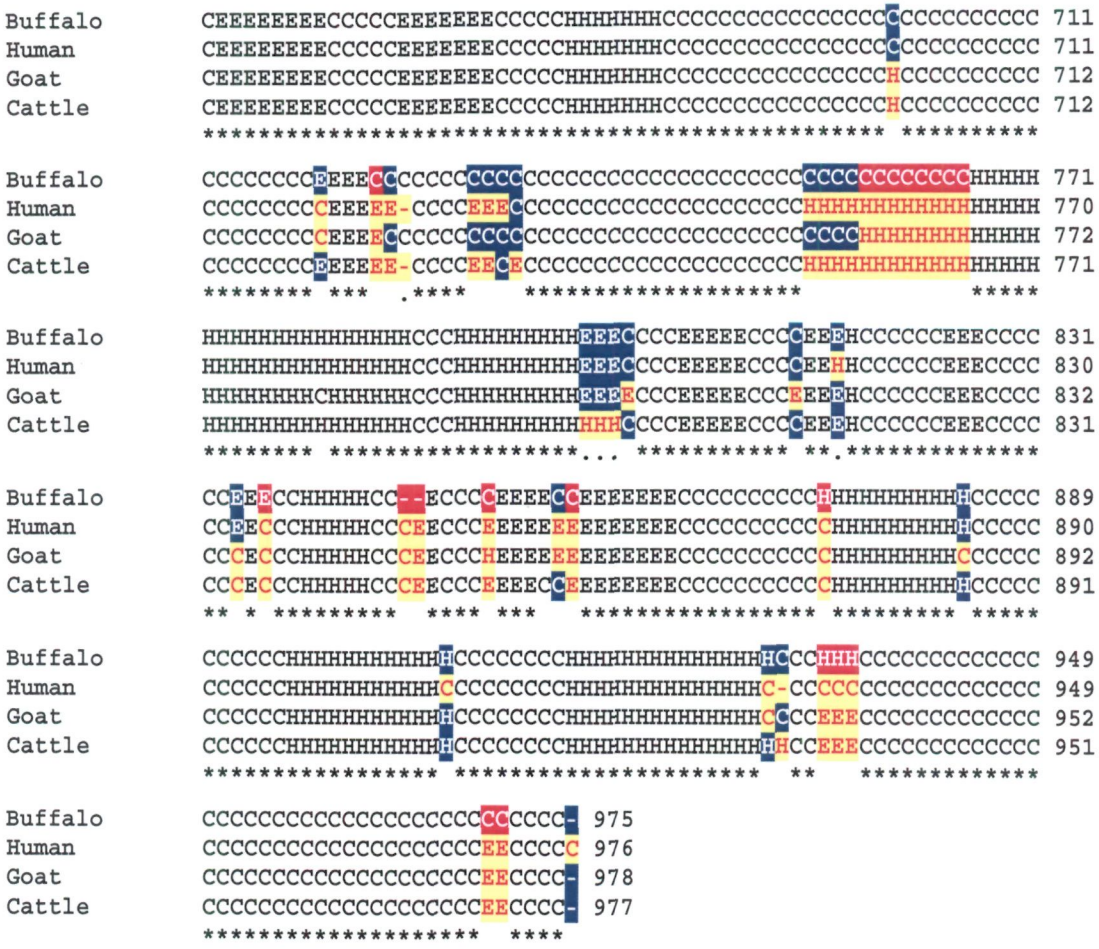


Figure 47. Multiple alignments of predicted secondary structures of c-kit from different species. Note the replacement of helix formed by 12 residues to a coil at one place in tyrosine kinase domain along with minor alterations throughout the protein.

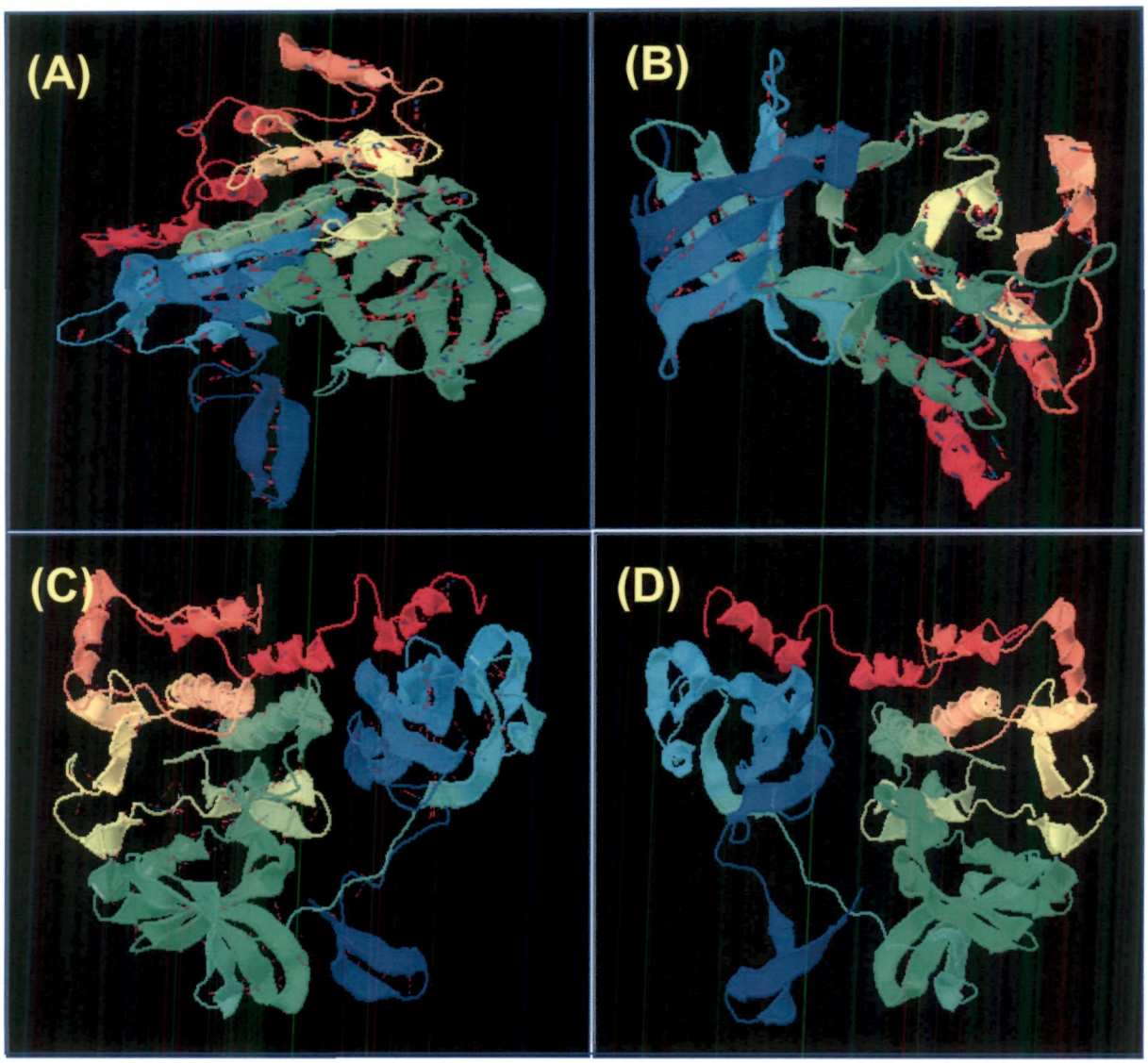


Figure 48. Multiple alignments of predicted tertiary structures of c-kit protein from different species. Buffalo (A), Cattle (B), Human (C) and Chimpanzee (D).

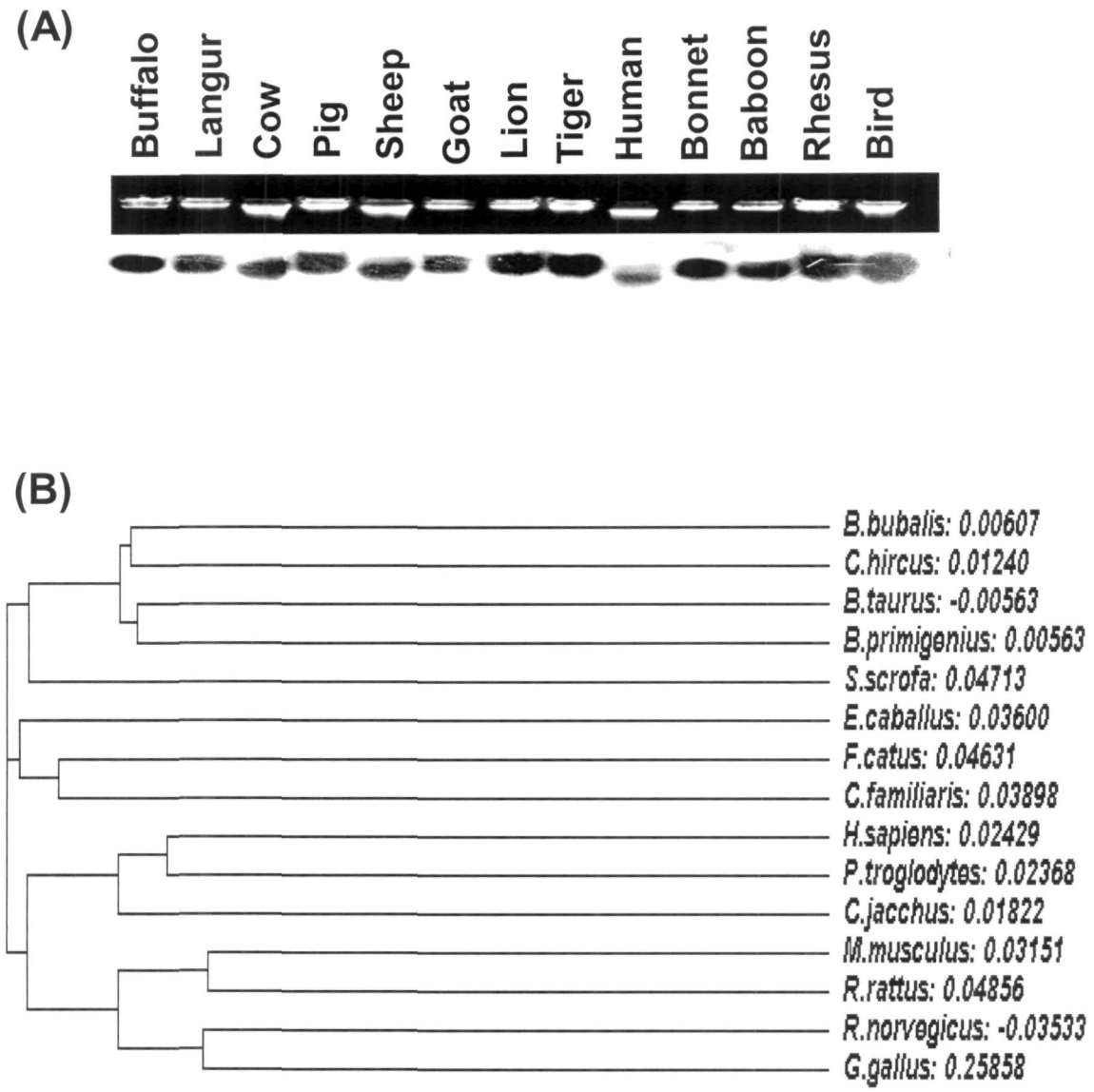


Figure 49. Evolutionary conservation of *c-kit* gene across the species based on the cross hybridization of buffalo *c-kit* with genomic DNA from different sources **(A)** and phylogenetic tree based on the sequence alignment using *c-kit* cDNA from different species **(B)**.

Table 15: Sequences of *c-kit* from fifteen different species with their accession numbers, nucleotide length, amino acid residues, no. of exons, chromosomal position and homology status with the buffalo *c-kit*

S. No.	Species	Accession numbers	Nucleotide length	Exon numbers	Translation length	Chromo-somal location	Homology with the buffalo c-kit
1.	*Bubalus bubalis*	DQ314491	2973 bp	21	975	NA	100%
2.	*Bos primigenius*	D16680	3069 bp	21	977	NA	98%
3.	*Capra hircus*	D45168	3771 bp	21	978	NA	98%
4.	*Bos taurus*	AF263827	2176 bp	21	724	6	97%
5.	*Equus caballus*	AF055037	2921 bp	-	939	3	89%
6.	*Homo sapiens*	XO6182	5084 bp	21	976	4q11-q12	86%
7.	*Mus musculus*	AY536430	2960 bp	21	979	5 42.0 cM	81%
8.	*Rattus rattus*	X62491	3705 bp	21	960	14	81%
9.	*Rattus norvegicus*	D12524	3816 bp	21	978	14	88%
10.	*Gallus gallus*	D13225	5045 bp	22	960	4	65%
11.	*Canis familiaris*	AY296484	3004 bp	20	979	13	88%
12.	*Felis catus*	NM_001009837	4222 bp	20	978	NA	88%
13.	*Sus scrofa*	AJ223229	2960 bp	NA	964	8p12	89%
14.	*Callithrix jacchus*	AB097502	2919 bp	NA	972	NA	86%
15.	*Pan troglodytes*	XM_517285	2919 bp	21	972	4	83%

4.2.1.7 *Buffalo genome has single copy of c-kit gene*

The copy number of the *c-kit* gene was estimated using Real Time PCR and absolute quantitation assays. The methodologies have been explained earlier in the section 3.10. Extrapolation of the standard curve showed a single copy of the *c-kit* gene per haploid genome of buffalo (Figure 50).

4.2.1.8 *Relative expression of c-kit receptor gene based on* Real Time *PCR*

Expression of *c-kit* gene was studied using cDNA templates from different tissues in a relative quantitation assays with the Real Time PCR (Figure 51A-B). Using β-actin as an internal control and liver cDNA as calibrator tissue, highest level of expression (163 folds) of *c-kit* was observed in testis (Figure 51C). This was substantiated by expression data from testis and ovaries of four different male and female animals, respectively, which showed 137-177 folds higher expression of *c-kit* in testis (Figure 51C-D). The *c-kit* expression in testis was found to be approximately 5 folds higher compared to that in ovaries. The details on the calculation with respect to level of expression have been described earlier.

In an independent assay, the same amount of testis and semen cDNA showed a Ct difference of 3.4 with the same set of primers (Table 16). This translated into ten fold higher expression of *c-kit* in testis as compared to that in the spermatozoa. The expression analysis was corroborated by expression data observed from ten different animals. However, this study was based on the relative quantitation of mRNA transcripts which may or may not translate into same amount of c-kit protein.

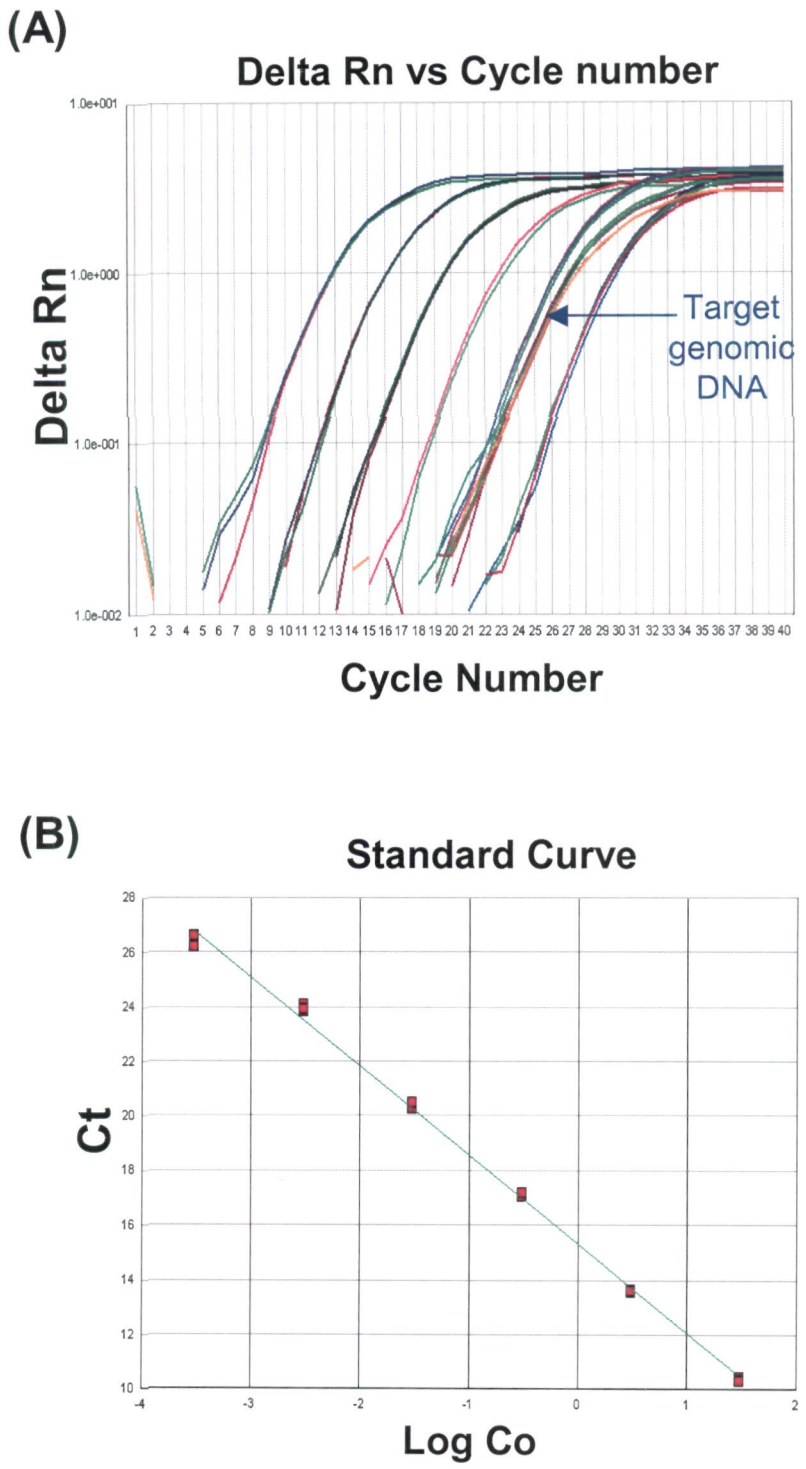


Figure 50. Copy number calculation for c-kit. Real Time PCR amplification plot based using 10 fold dilution series of recombinant plasmid containing *c-kit* gene **(A)** and standard curve **(B)** based on this dilution series which lead to the copy number status of c-kit as '1'.

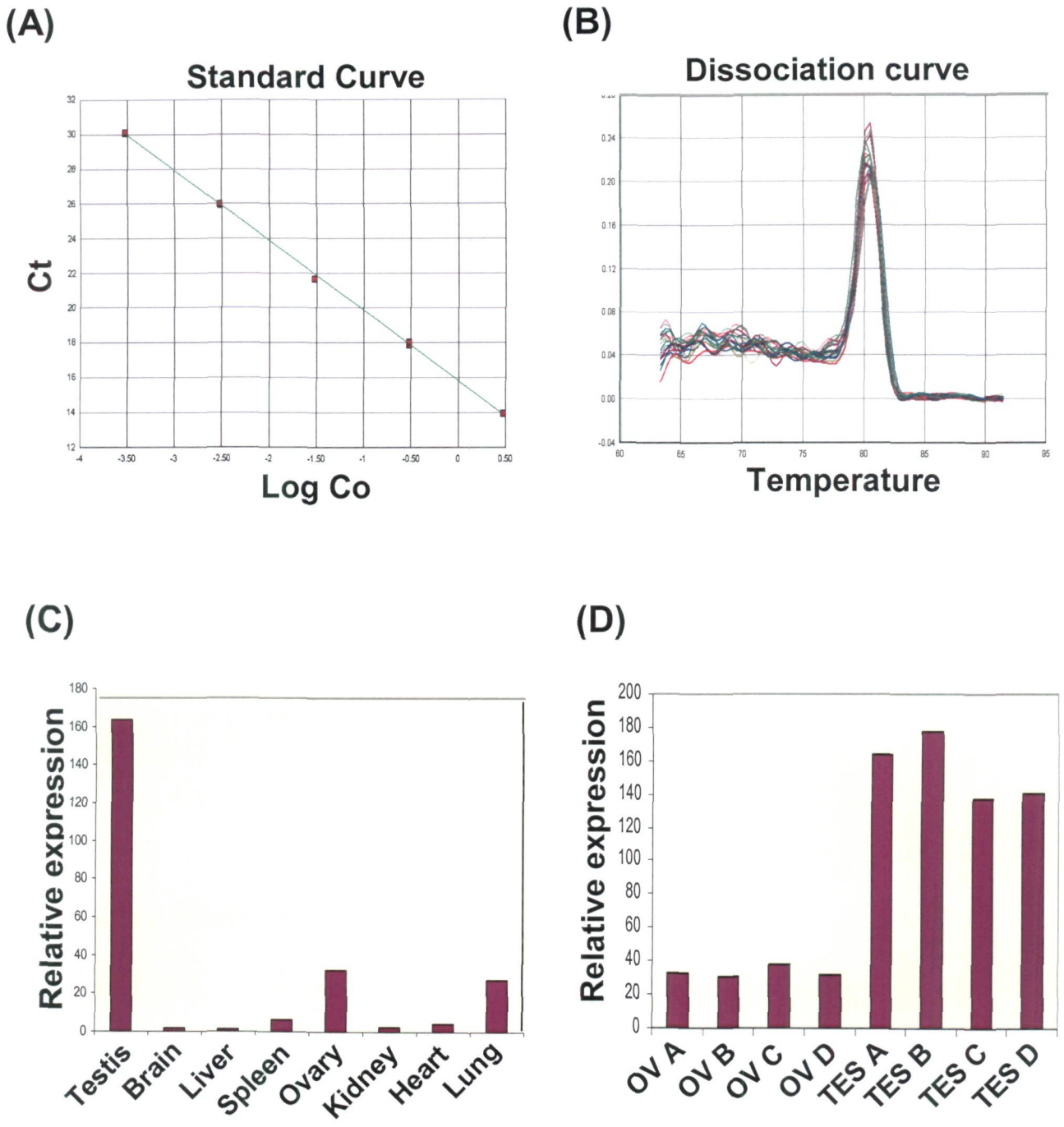


Figure 51. Standard **(A)** and Dissociation Curves **(B)** based on Real Time PCR using SYBR Green Dye and five fold dilution series of the cDNA. Single peak in the dissociation curve reflects high specificity of the primer. Real Time PCR based relative quantitation of *c-kit* expression in different tissues of buffalo showing maximum expression (163 folds) in testis compared to that in liver taken as an endogenous control **(C)** and amplification of the same in testes and ovaries of four different animals **(D)**. Note the expression of *c-kit* in testis ranging from 137-177 folds higher than that in any other tissue.

Table 16: Mean Ct values for *c-kit* in testis and semen samples using same amount of cDNA from both ΔCt of 3.3 (approx) demonstrates 10 fold difference in the expression levels.

S.N.	Testis		Semen		ΔCt
	Sample ID	Ct	Sample ID	Ct	(Semen-Testis)
1	tesA	24.303	255s	27.603	3.30
2	tesB	23.680	256s	27.020	3.340
3	tesC	24.00	257s	27.326	3.326
4	tesD	24.330	326s	27.630	3.30

4.2.2 Secreted modular calcium binding factor-1 (*Smoc*-1)

4.2.2.1 *Isolation and cloning of buffalo Smoc-1*

The cloning strategy for the isolation of the full length *Smoc*-1 CDS of 3474 bp and 1933 bp has been demonstrated in the figure 52A-B and described earlier in the section 3.5. Briefly, all the fragments were PCR amplified (Figure 53) using different primer pair sets (Table 3), and were subsequently cloned (Figure 54) and sequenced.

Clone I contained an insert of 1263 bp (PSmoc-1) lacking 5'/3' UTRs and a signal peptide sequence. Clone II (1414 bp) covered complete coding sequence from nucleotides 119-1603 but with partial 5' & 3' UTRs. The 3' UTR was covered by three fragments represented by clone III, IV & V of which clone III covered nucleotides from 1461-2473; clone IV, 2307-3328; and clone V, 2435-3428 (Figure 52). The polyadenylation signals were accessed with 3' RACE followed by sequencing of 30 recombinants. This resulted in the identification of two other clones (Clone VII & VIII). Clone VII represented nucleotides 1407-1915 followed by 17mer Poly(A) tail. Clone VIII covered 2435-3474 encompassing 18mer Poly(A) tail. The 5' UTR represented by clone VI (nucleotides 1-649) was generated by 5' RACE (Figure 53 and 54).

Following this strategy, the full length CDS of *Smoc*-1 was deduced from the different overlapping fragments (Figure 52 and 55), and then submitted to the GenBank and the accession numbers were obtained (GenBank accession numbers: DQ159955 and EF446167). The GC rich 5'-UTR of 239 bp was followed by an initiating ATG codon and the terminating TAA codon fell at nucleotide 1545 (Figure 55). Thus, translation of the sequence from nucleotide 240 to 1544 encodes a putative protein of 435 amino acids with a calculated molecular mass of 48332 Da with a predicted oxidoreductase activity. Multiple sequence alignments showed that buffalo Smoc-1 protein is 95% and 98% identical to the human & cattle *Smoc*-1, respectively (Table 17).

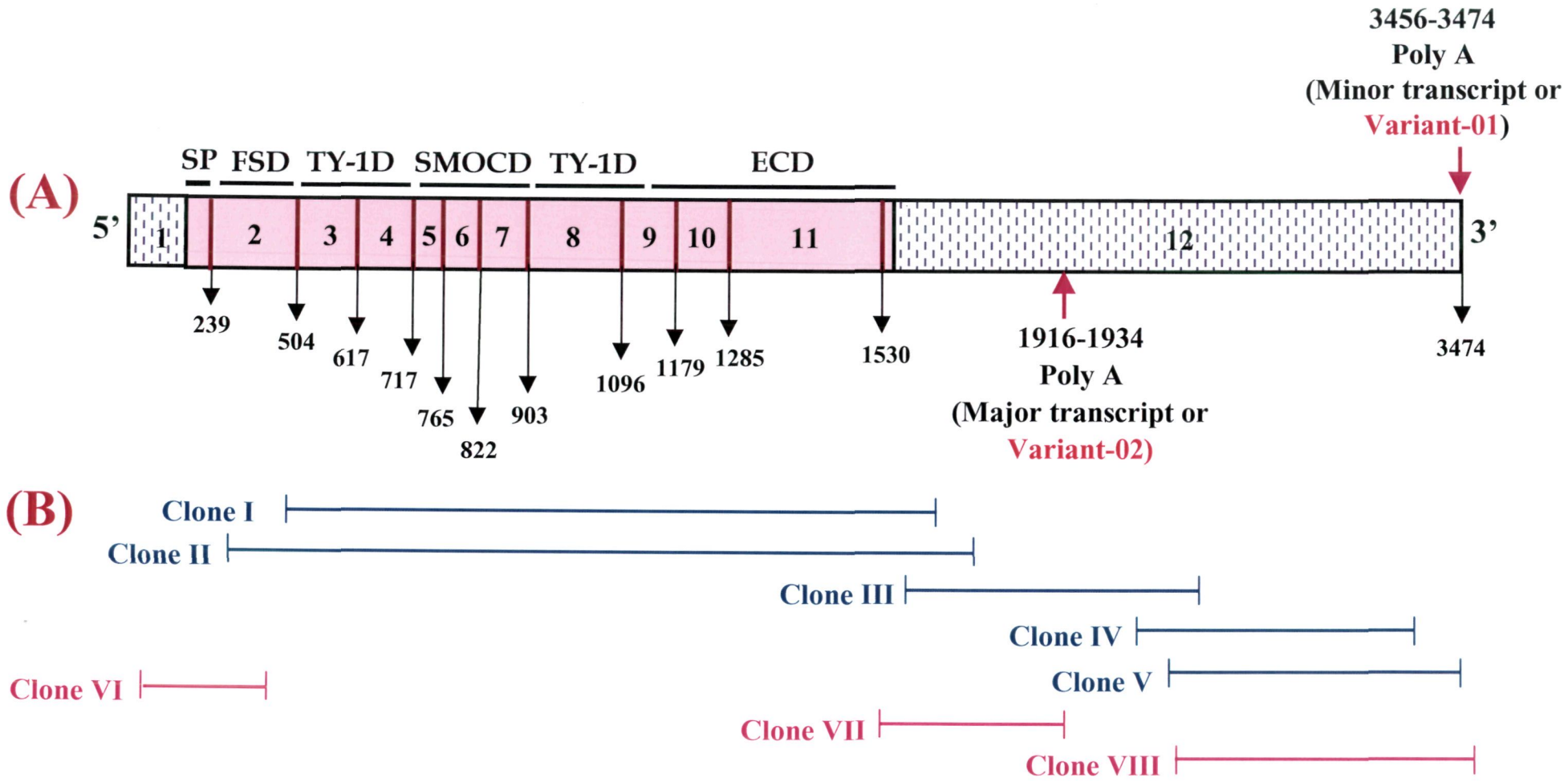


Figure 52. Diagrammatic illustration showing cloning strategy of buffalo *Smoc*-1. *Smoc*-1 structure representing 5'/3'UTRs, domain organization and nucleotide boundary of each exon is shown in **(A).** The strategy for isolation of the *Smoc*-1 is given in **(B).** Different fragments generated by end point PCR (blue) and RACE (pink) used to deduce the full length CDS are shown along with their nucleotide boundaries. Clone I covered nucleotides 318-1580; clone II, 119-1603; clone III, 1461-2473; clone IV, 2307-3328; and clone V, 2435-3428; clone VI, 1-649; clone VII, 1407-1915 and clone VIII, 2435-3474. Two transcript variants of *Smoc*-1 with their 3'UTR length variation are shown. Poly(A) tails for both the variants, -01 (3474 bp) and -02 (1934 bp) are marked by arrows in 'A'.

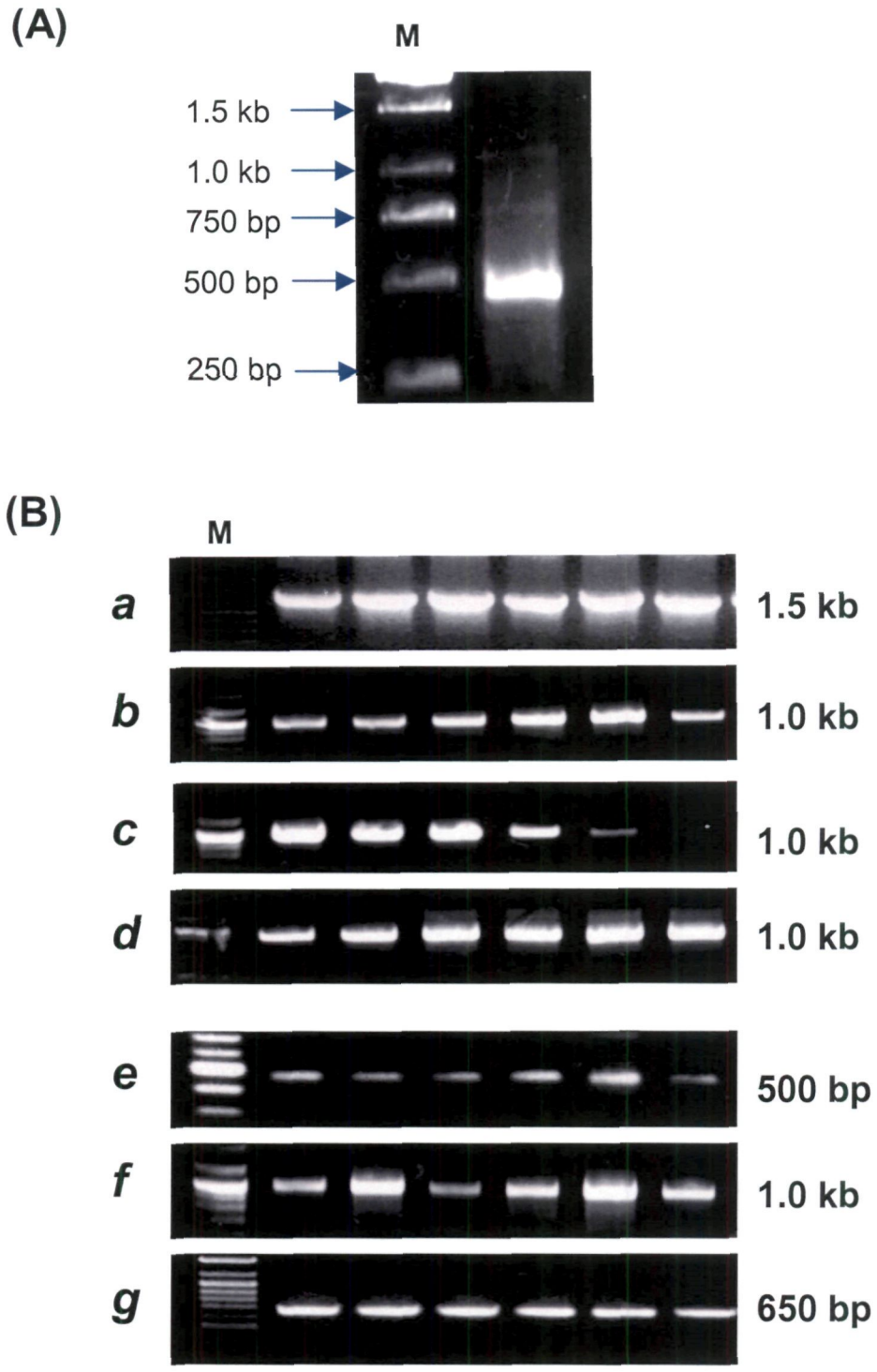


Figure 53. Isolation of *Smoc*-1 gene by RACE **(A)** and PCR amplification using different primer sets designed **(B).** The amplification product sizes are given on the right of each gel picture. This product was subsequently cloned and sequenced.

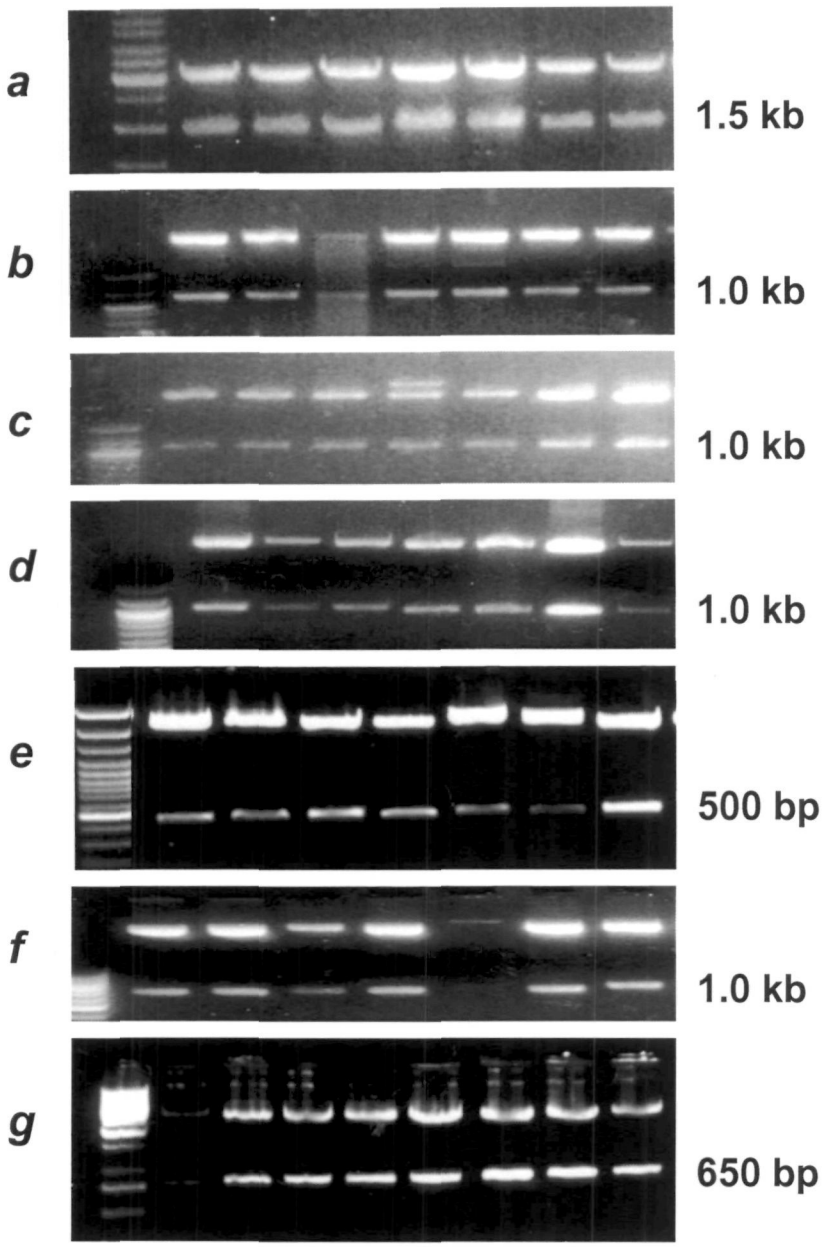


Figure 54. Representative pictures for the restriction analyses of the recombinant clones of Smoc-1: using *EcoR*1 **(a-g).** The sizes of insert released for particular fragment has been shown in the right of each gel picture. The molecular size marker "M" is given in base pairs. .

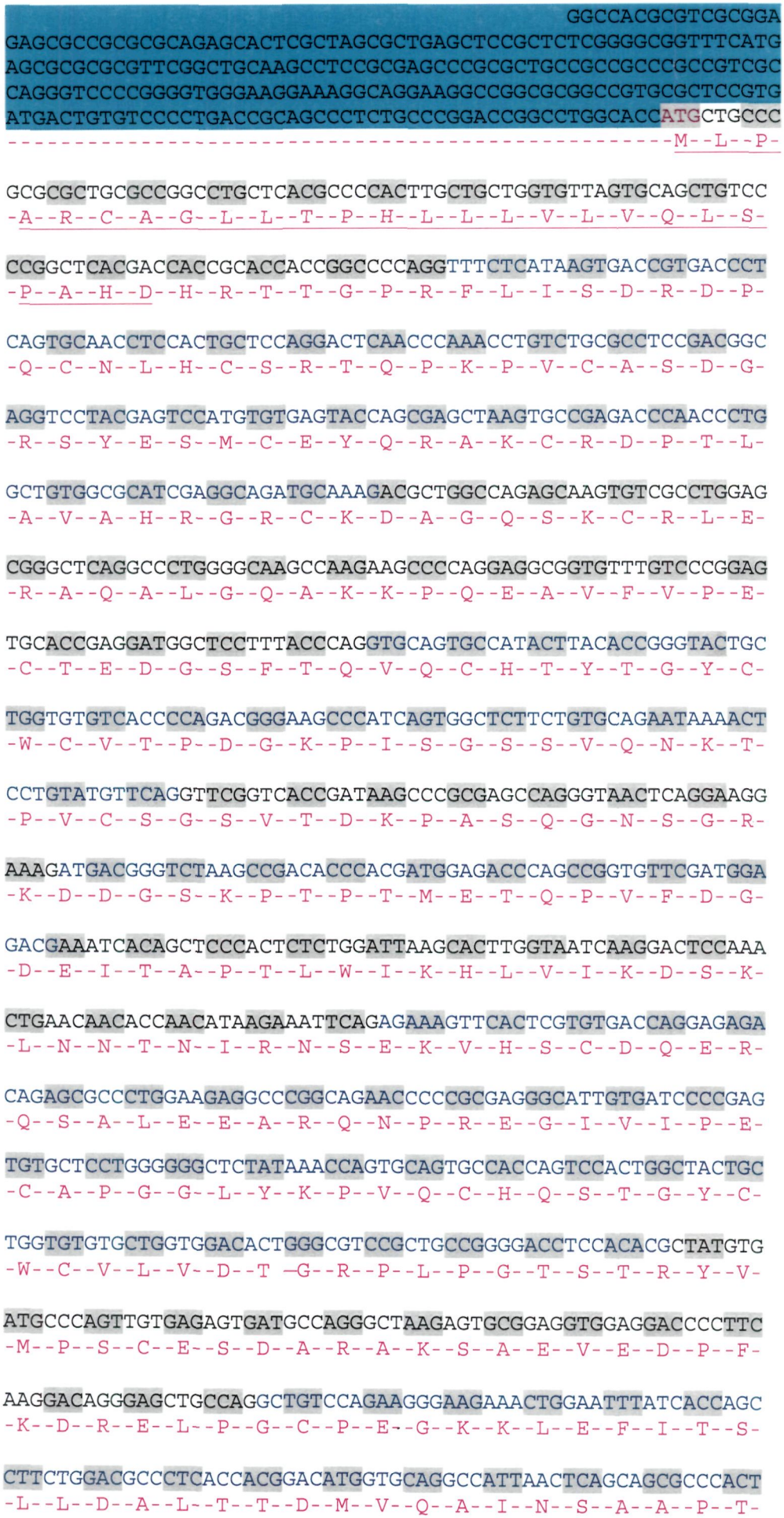


Figure 55

Contd/-

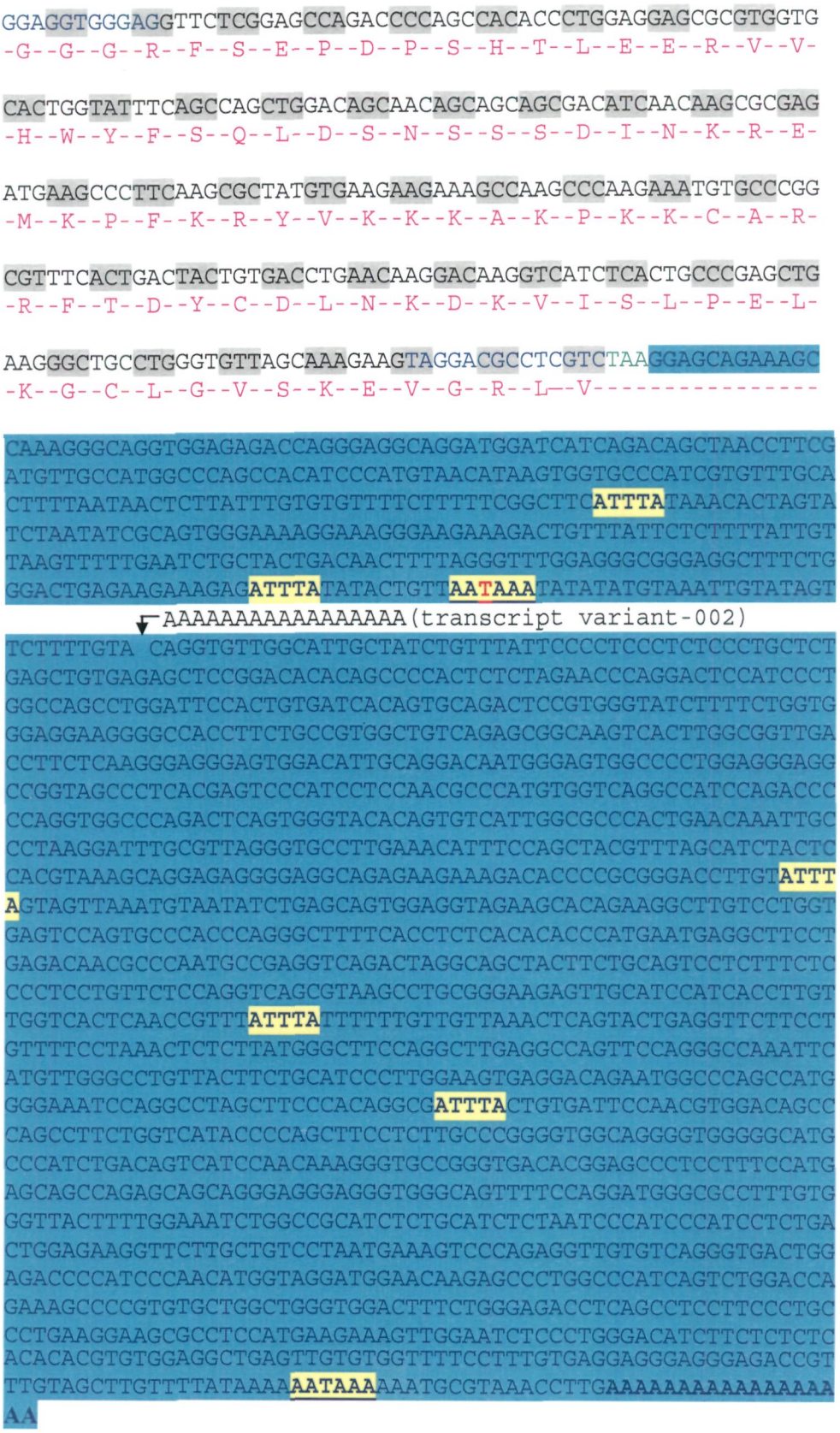


Figure 55. Complete cDNA sequence with the deduced amino acid sequences. The exons are shown in alternate colors and the start/stop codons in boldface & violet. The putative signal peptide sequence is underlined. The RNA instability motifs (ATTTA) are overshadowed yellow whereas polyadenylation signals (AATAAA) are underlined and shadowed yellow. The Poly(A) tail for transcript variant-02 is indicated by an inserted arrow and for variant-01 in boldface at the end of the sequence. Note the sequence 5'AAAAAA3' in transcript variant-01 is replaced by 5'AATAAA3' in variant-02 as evident by sequence analysis of 25 recombinant clones.

Table 17: Secreted modular calcium binding protein-1 (*Smoc*-1) from different species and their homology with that of water buffalo *Bubalus bubalis*#

S.N.	Species	Accession numbers (Ensembl/NCBI)	Transcript length	Full length gene (In Kb)	No. of exons	Amino acid residues	Chromo-somal location	Homology with buffalo *Smoc*-1		
								CDS	Amino acids	Secondary structure
1.	*Bubalus bubalis*	DQ159955	3274	NA	12	435	11	100%	100%	100%
2.	*Homo sapiens*	ENSG0000198732/ AJ249900	3666	172.94	12	434	14q24.2	91%	95%	93%
3.	*Pan traglodytes*	ENSPTRT0000011881/ XM_510036	3,669	154.09	14	435	14	90%	94%	94%
4.	*Bos taurus*	XM_612029	3473	NA	NA	434	10	98%	98%	97%
	Mus musculus	ENSMUSG0000021136/NM_022316	3472	179.58	13	463	12d3	84%	93%	89%
6.	*Rattus norvegicus*	ENSRNOG00000005998/ NM_0102835	1359	193.42	12	452	6q24	85%	92%	89%
7.	*Canis familiaris*	ENSCAFT0000026288	1452	152.3	12	459	8	87%	95%	85%
8.	*Gallus gallus*	ENSGALG00000009415	1404	98.76	12	468	5	78%	84%	83%

The detailed information on *Smoc*-1 including accession numbers, gene length, exon numbers and chromosomal location are given.

4.2.2.2 *Buffalo Smoc-1 shows two transcript variants*

Northern blot detected two bands of 3.5 kb and 2.0 kb (Figure 56) which were confirmed to be two transcript variants of *Smoc*-1 with RACE and sequencing (Figure 52 and 55), variant-01 of 3474 bp (GenBank: DQ159955) and variant-02 of 1933 bp (GenBank: EF446167). Both the variants encoded for identical proteins but differed at their 3'UTR length, polyadenylation signals and Poly(A) tails. In the 3' UTR of variant-01 & -02, five & two copies of mRNA instability motif (ATTTA), respectively, were observed. In addition, there were orthodox polyadenylation signals (AATAAA), 1787 and 334 bp downstream of the translation termination codon, in transcript variant-01 and -02, respectively (Figure 55). Interestingly, two types of transcripts of *Smoc*-1 have been reported independently in the literature for human (GenBank: AJ249900 and BC011548) and cattle (GenBank: XM_612029 and NM_01079771). Database search and multiple nucleotide sequence alignment of both the variants showed their sequence conservation across various species even for the presence of poly(A) tails (Figure 57).

4.2.2.3 *Structure of the buffalo Smoc-1 and its phylogenetic delineation*

The buffalo *Smoc*-1 gene was found to have 12 exons, varying in length from 48 to 1916 bp in variant-01 and 48 to 402 bp in variant-02 (5^{th} being the smallest and 12^{th}, the longest one) (Figure 52 and 55). Each domain of Smoc-1 is encoded by one or more exons and the domain borders coincide with the conserved splice sites. The buffalo *Smoc*-1 cross-hybridized to genomic DNA from 13 different species with almost equal signal intensity, confirming its faithful conservation across the species (Figure 58A). The *Smoc*-1 from different mammals was studied *In-silico* for the transcript/gene and translation lengths, number of exons, chromosomal position, and the homology with that from the buffalo (Table 17).

Thereafter the phylogenetic analyses demonstrated that the buffalo *Smoc*-1 is closer with that of cattle followed by in human and chimpanzee, whereas dog was found to be the most distant one (Figure 58B). In

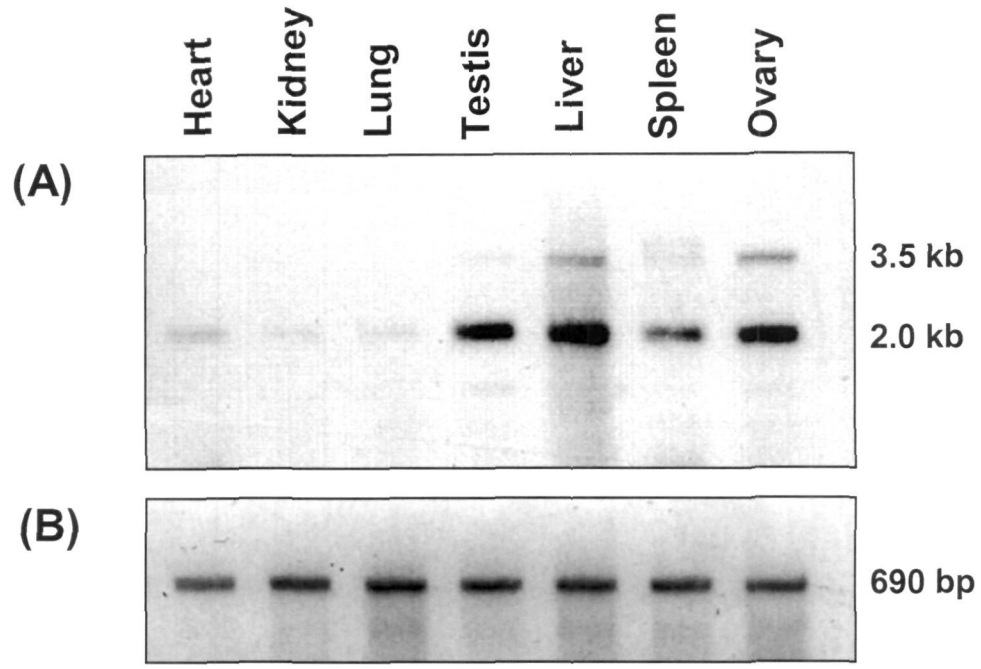


Figure 56. Two transcript variants of *Smoc*-1. Northern blot showing two transcript variants of *Smoc*-1 in different somatic and gonadal tissues of water buffalo. Note two distinct bands with varying intensity in each tissue along with highest expression in liver, and lowest in lung, kidney, and hear **(A)** and the ß-actin gave equal signal in each lane **(B)**.

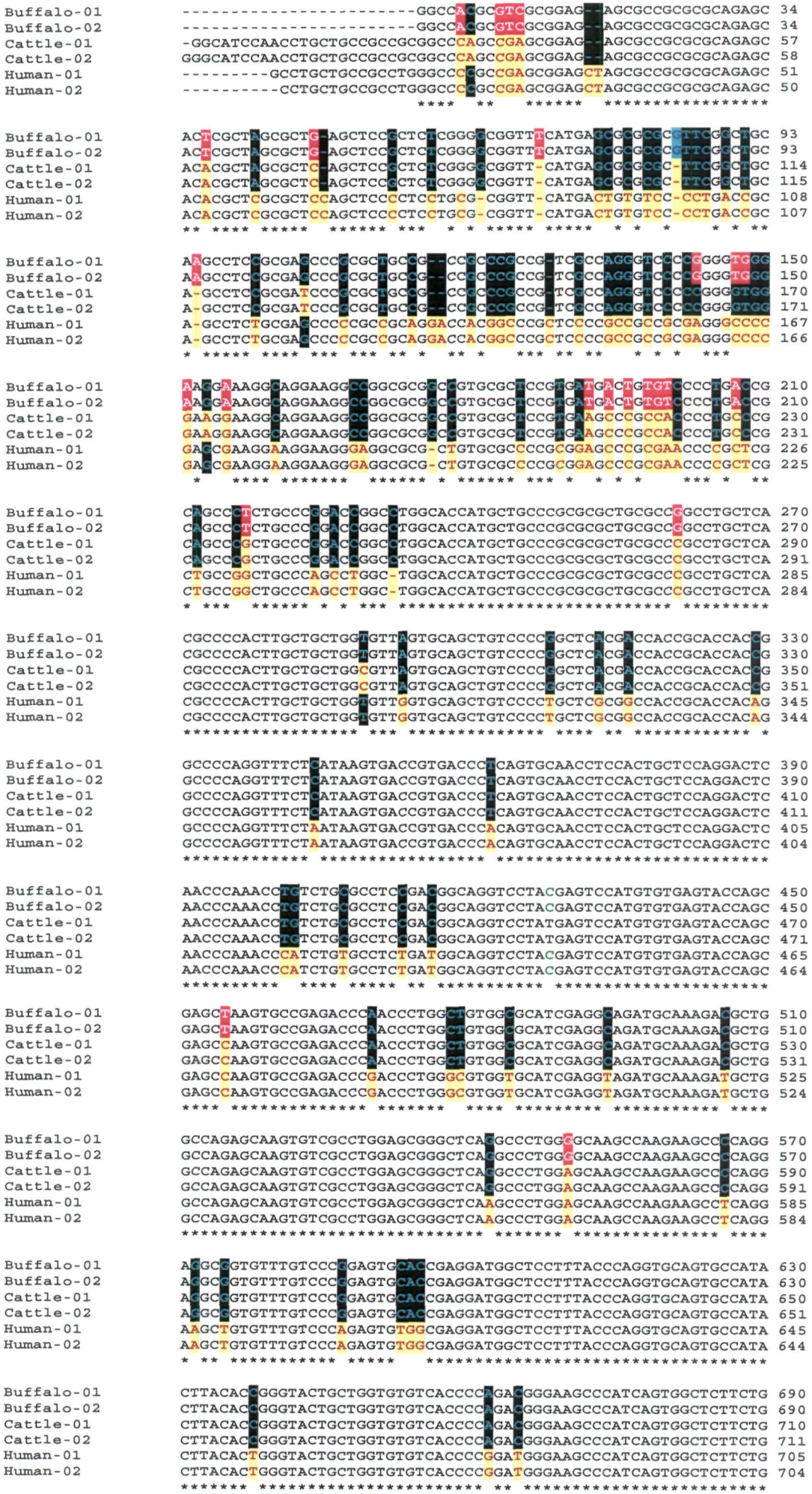


Figure 57

Contd/-

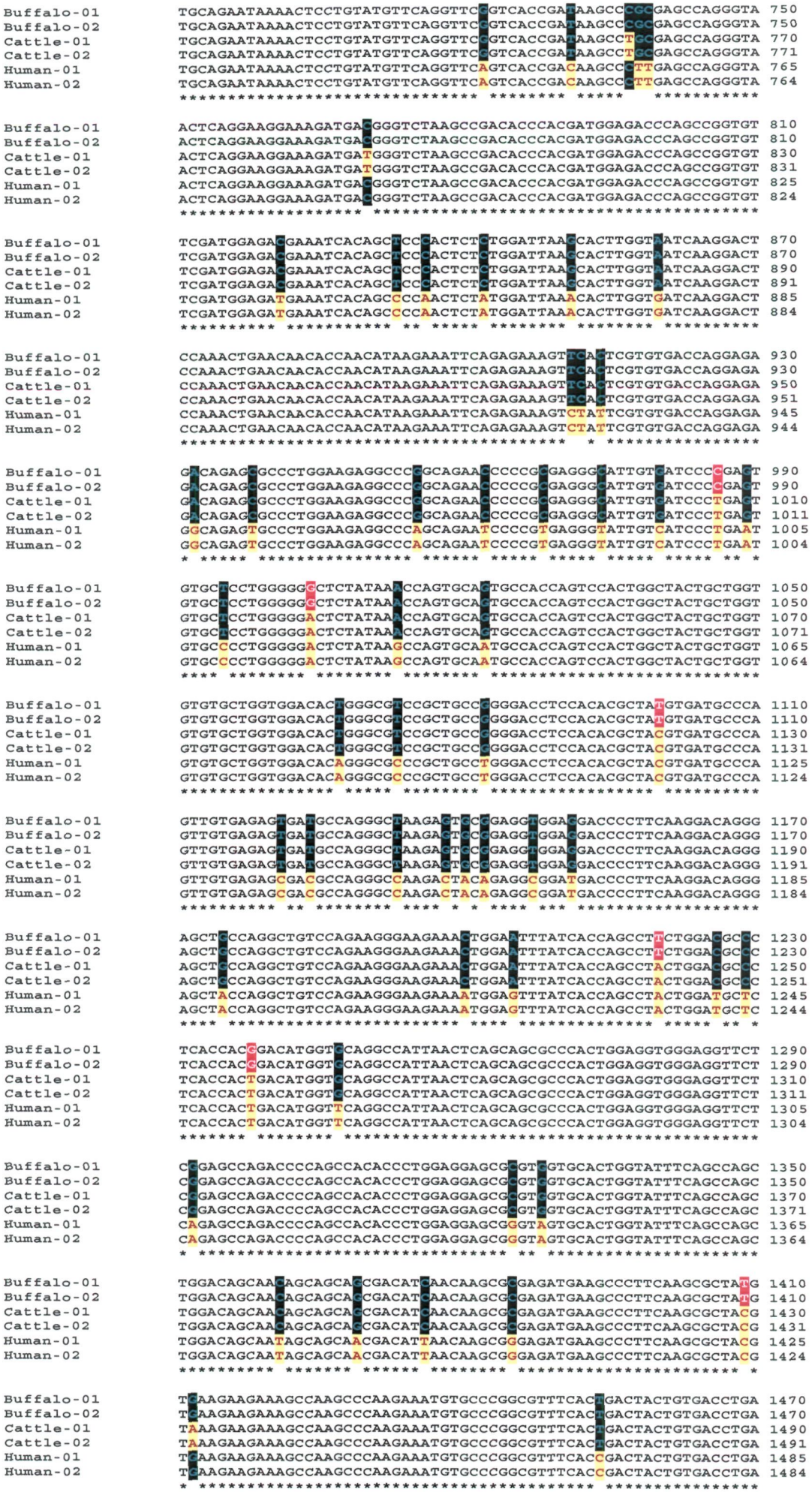


Figure 57

Contd/-

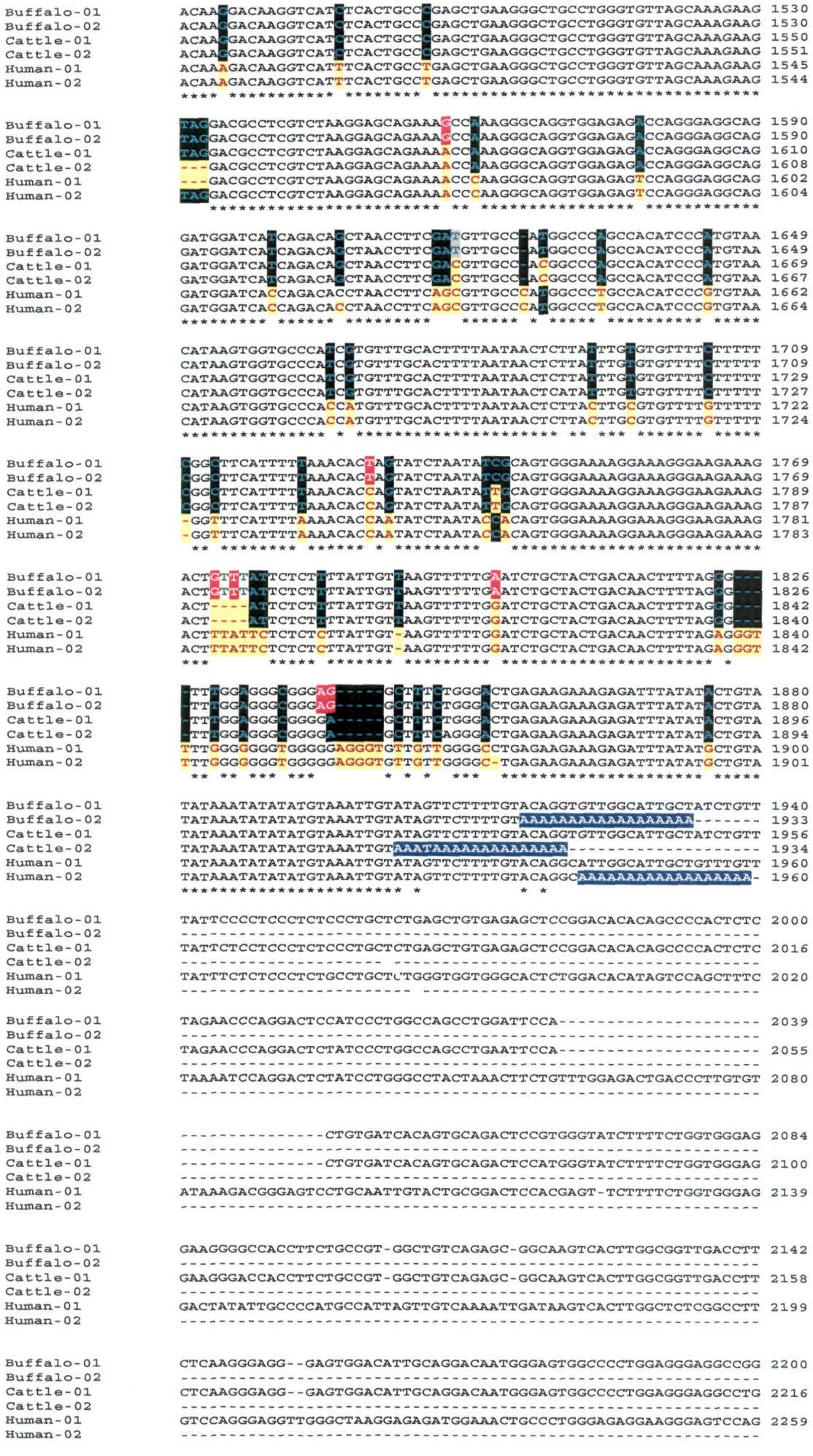


Figure 57

Contd/-

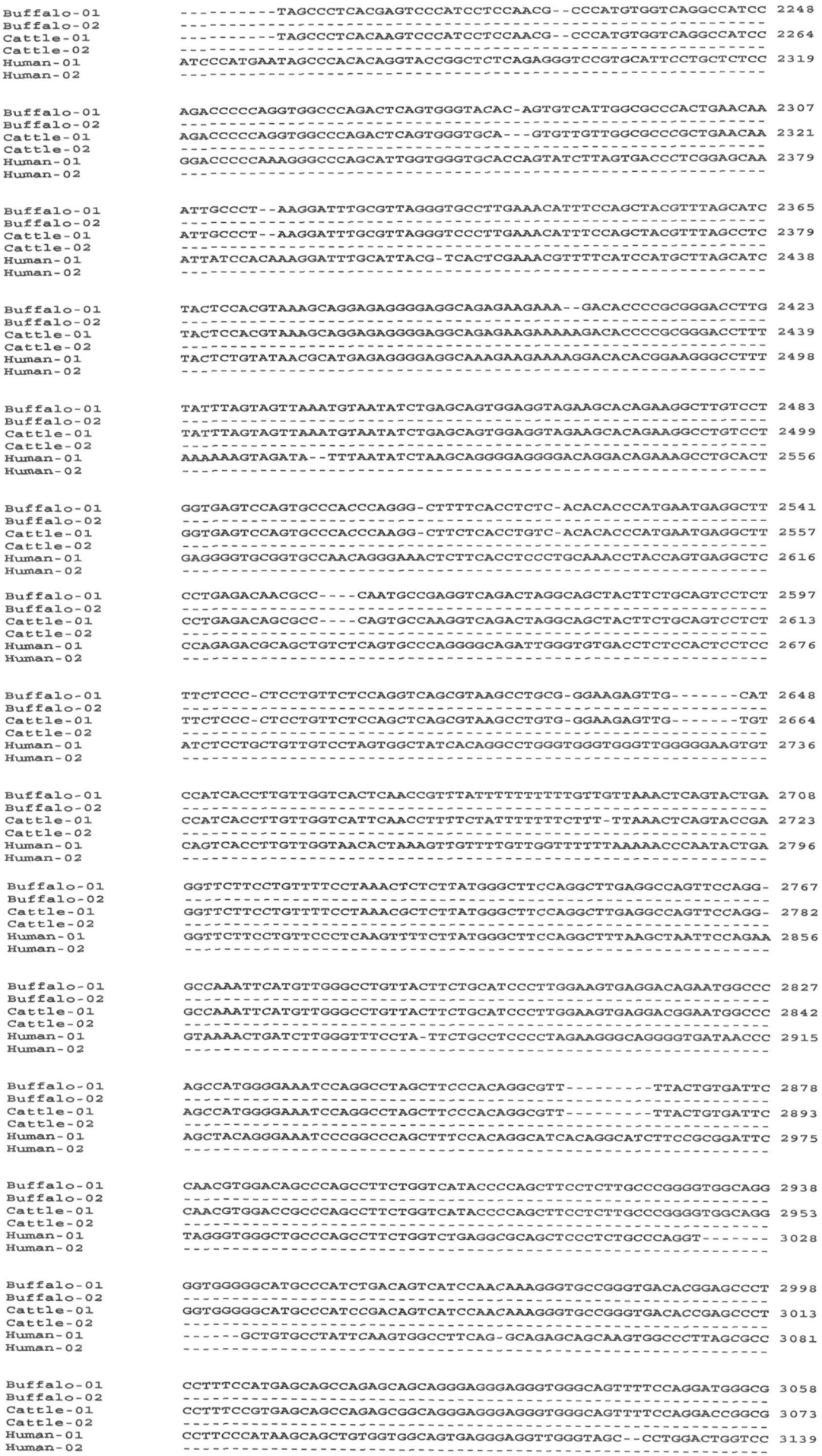


Figure 57

Contd/-

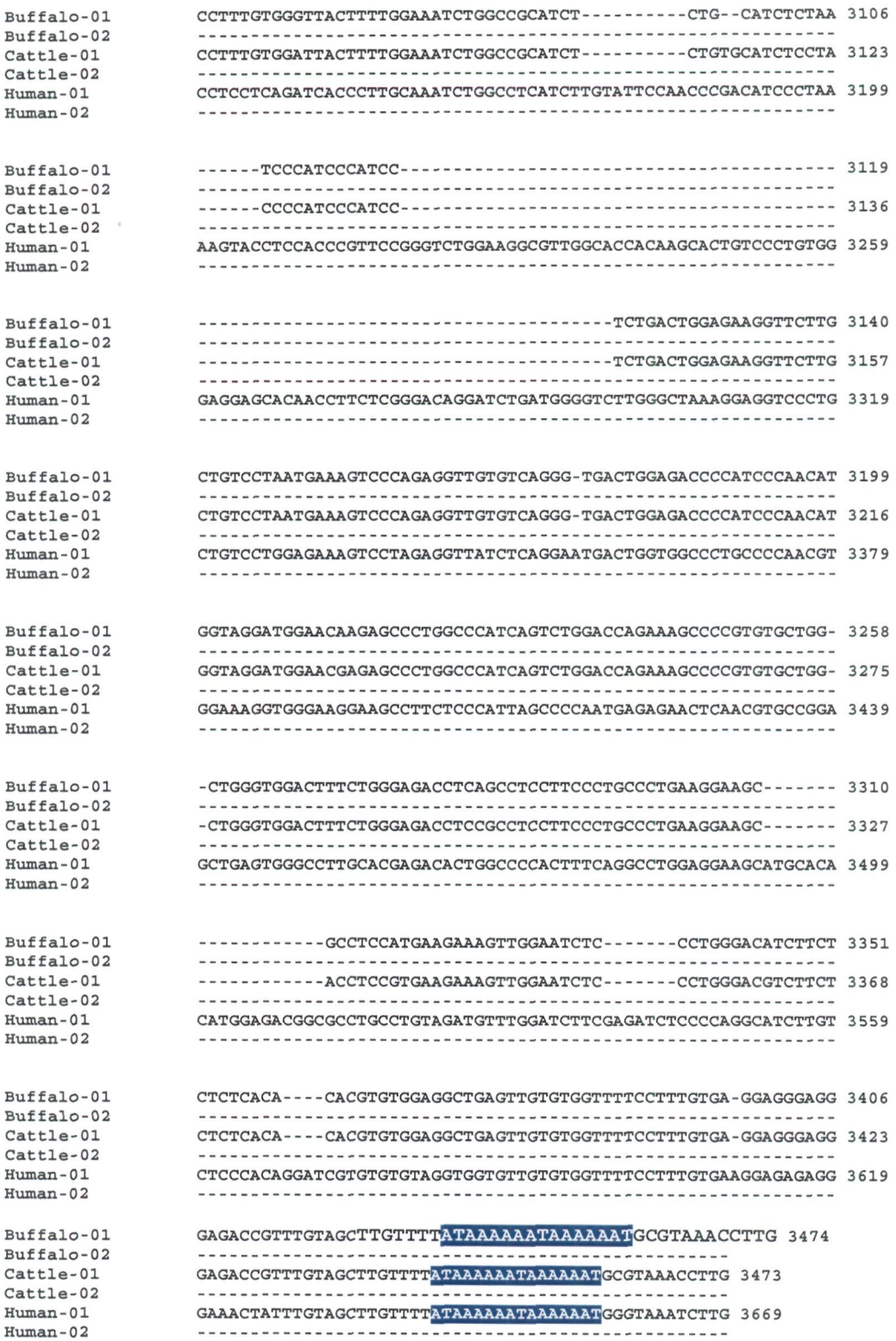


Figure 57. Multiple nucleotide sequence alignment for both transcript variants from buffalo, cattle and human. In these species, variant-02 is shorter and almost of the same size due to possible conserved splice site in the 3'UTR. Polyadenylation signals (bold face) and Poly(A) tails (bold face and blue) are conserved in each variant in all the species. The nucleotides unique to buffalo are overshadowed red. The changes specific to buffalo and cattle are shown in black background and the ones similar to human are in yellow background.

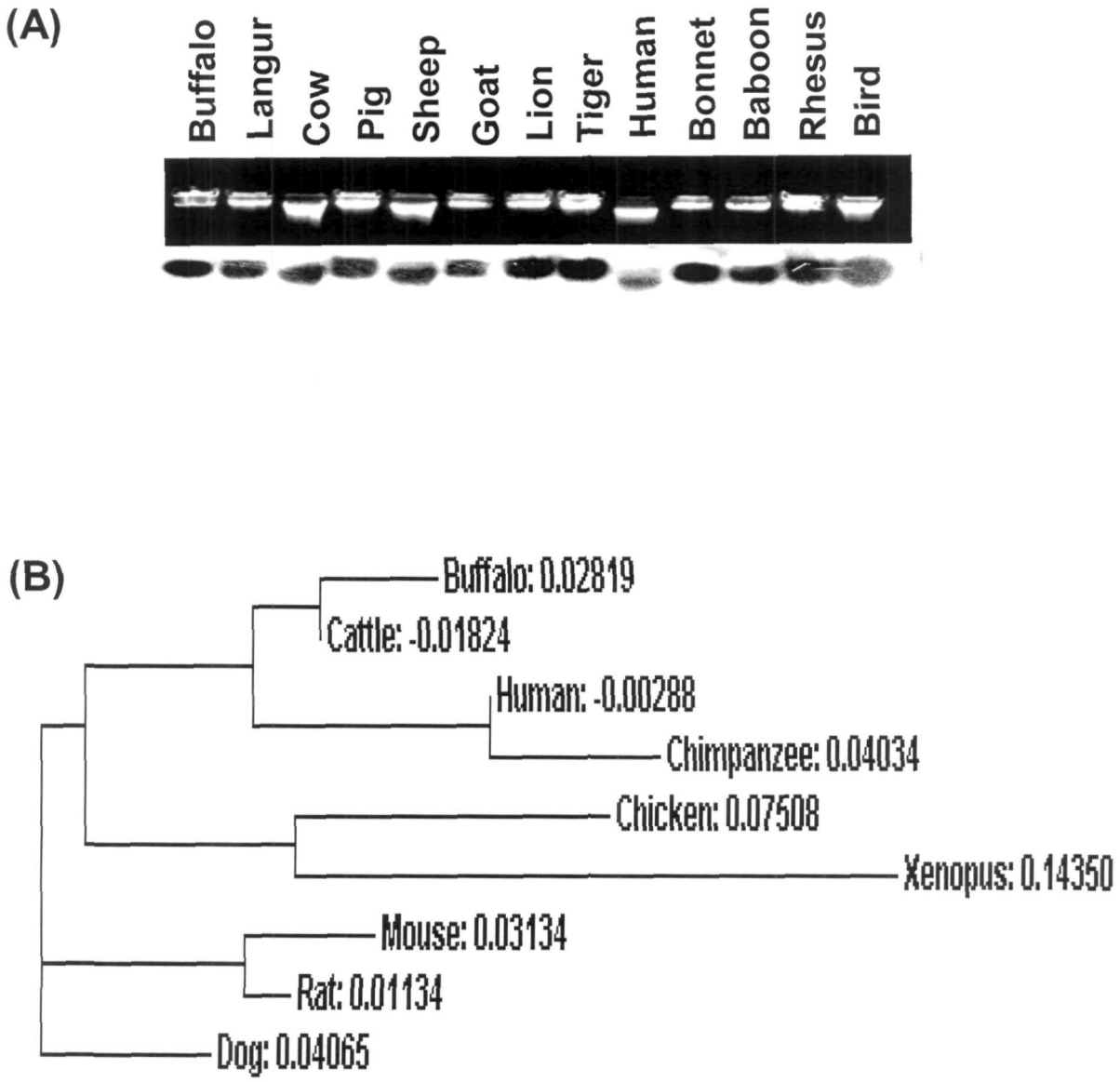


Figure 58. Evolutionary conservation of *Smoc*-1 gene across the species. Cross hybridization of buffalo *Smoc*-1 with genomic DNA from different species **(A)**, Phylogenetic tree based on sequence alignment of *Smoc*-1 gene(s) from different species **(B)** and neighbor joining tree based on BLAST result showing homology across the species with their accession numbers **(C)**. Note that this gene is phylogenetically conserved across the species.

Contd/-

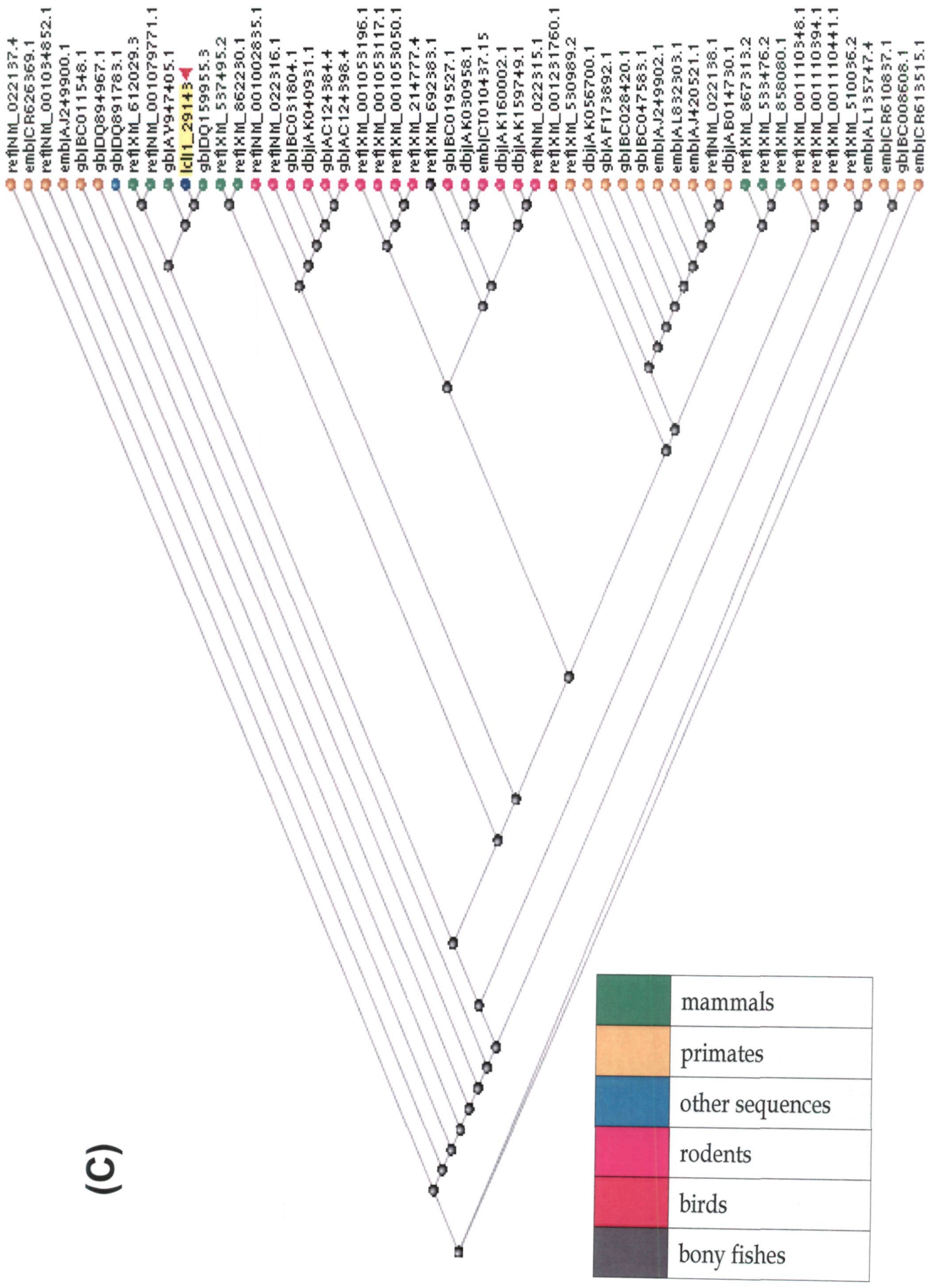

(C)
refNM_022137.4
embCR626369.1
refNM_001034852.1
embAJ249900.1
gbBC011548.1
gbDQ894967.1
gbDQ891783.1
refXM_612029.3
refNM_001079771.1
gbAY947405.1
lcl1_29143
gbDQ159955.3
refXM_537495.2
refXM_862230.1
refNM_001002835.1
refNM_022316.1
gbBC031804.1
dbjAK040931.1
gbAC124384.4
gbAC124398.4
refXM_001053196.1
refXM_001053117.1
refXM_001053050.1
refXM_214777.4
refXM_692383.1
gbBC019527.1
dbjAK030958.1
embCT010437.15
dbjAK160002.1
dbjAK159749.1
refNM_022315.1
refXM_001231760.1
refXM_530989.2
dbjAK056700.1
gbAF173892.1
gbBC028420.1
gbBC047583.1
embAJ249902.1
embAL832303.1
embAJ420521.1
refNM_022138.1
dbjAB014730.1
refXM_867313.2
refXM_533476.2
refXM_858080.1
refXM_001110348.1
refXM_001110394.1
refXM_001110441.1
refXM_510036.2
embAL135747.4
embCR610857.1
gbBC008608.1
embCR613515.1
mammals
primates
other sequences
rodents
birds
bony fishes

Figure 58

addition, the buffalo *Smoc*-1 also showed homology with other classes such as from the birds, rodents and bony fishes (Figure 58C). However, buffalo *Smoc*-1 showed point nucleotide alterations, insertions and deletions at several nucleotide positions in comparison to that in other species (Figure 57 and 59) leading to alterations in the encoding amino acid sequences (Figure 60). Briefly, exon 8 was found to be more diverse at nucleotide level whereas exon 2 comprising the follistatin domain was most divergent at amino acid level. Exon 6 was the most conserved across the species (Figure 59-60). Details of the *Smoc*-1 gene(s) from different species along with their accession numbers are given in the table 17.

4.2.2.4 *Domain Organization of Smoc-1*

In-silico studies of the buffalo Smoc-1 protein demonstrated presence of all the domains characteristic to the BM-40 family to which this gene belongs (Figure 61). Accordingly, first 26 amino acids at the N-terminus conform well to the signal peptide consensus (Figure 60 and 62) ending with a signal peptidase cleavage site (Kizawa *et al.*, 2005) and thus the mature Smoc-1 is comprised of 409 amino acids. Like the humans, all the essential features of each domain are conserved in buffalo. Since no transmembrane-spanning hydrophobic domain is present in the sequence, Smoc-1 is presumably secreted out from the cells. Further scrutiny allowed the distinction of five modules, an FS domain (Figure 61A), a TY domain (Figure 61B), a Smoc-1 unique domain (Figure 61C), a second TY domain (Figure 61D) and an EC domain (Figure 61E). Residues 42-88 are homologous to the canonical FS domain, composed of two sub-domains with the second being similar to the Kazal domain.

Structure-based alignment showed that all the six cysteines and the features of secondary structure are conserved in both the TY domains of buffalo *Smoc*-1. Further, two TY domains (residues 89-159 and 222-293) are separated by 62 amino acids unique to the Smoc-1. Detailed analysis unveiled a potential *N*-glycosylation site at Asn-214 and five O-glycosylation sites at Thr-155, -82, -184, -187 and -345 in buffalo Smoc-1. The C-terminus is homologous to the characteristic amphipathic α-helix and the helix-loop-helix motifs of the ECD of BM-40. However, based on

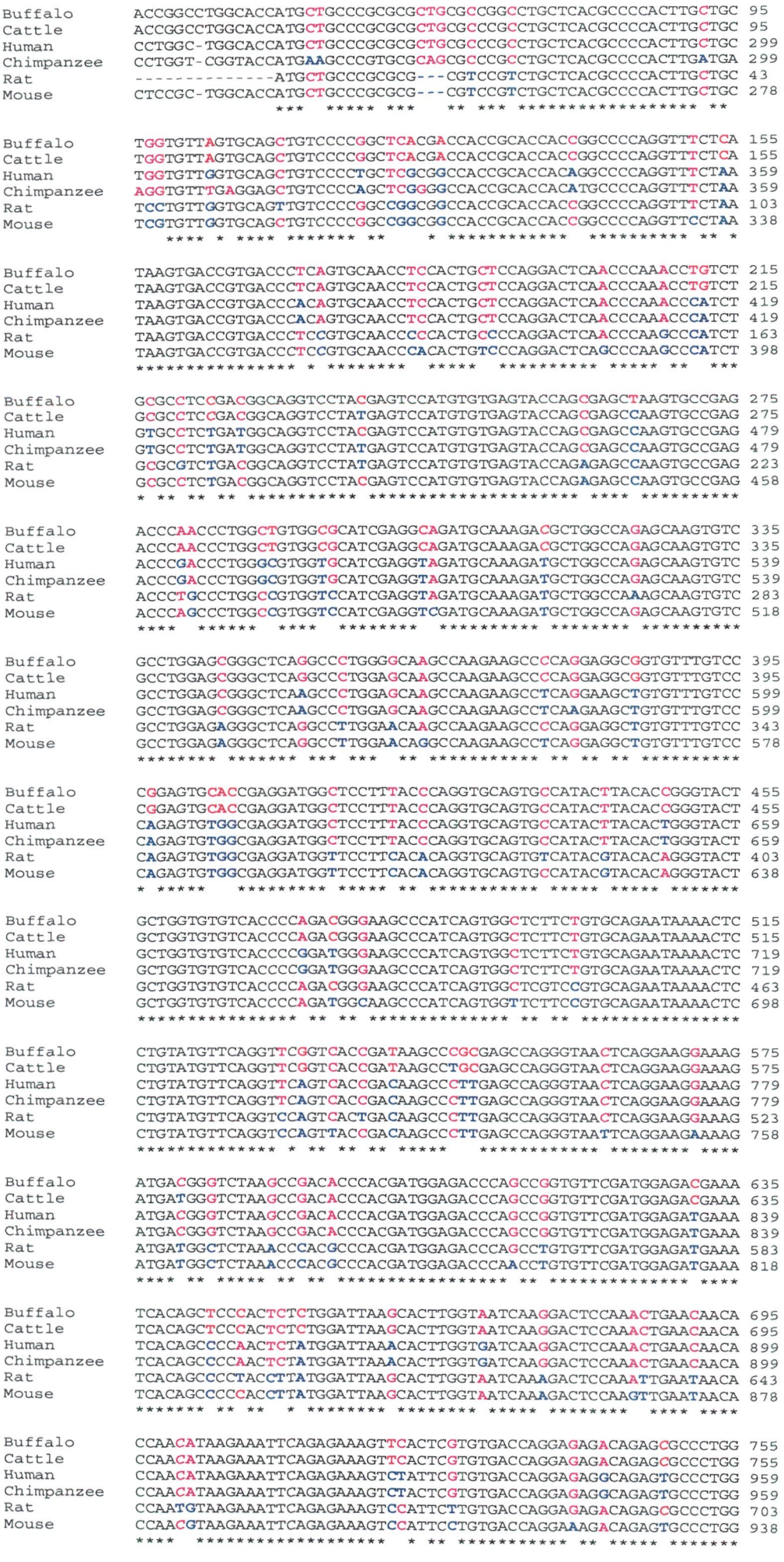


Figure 59

Contd/-

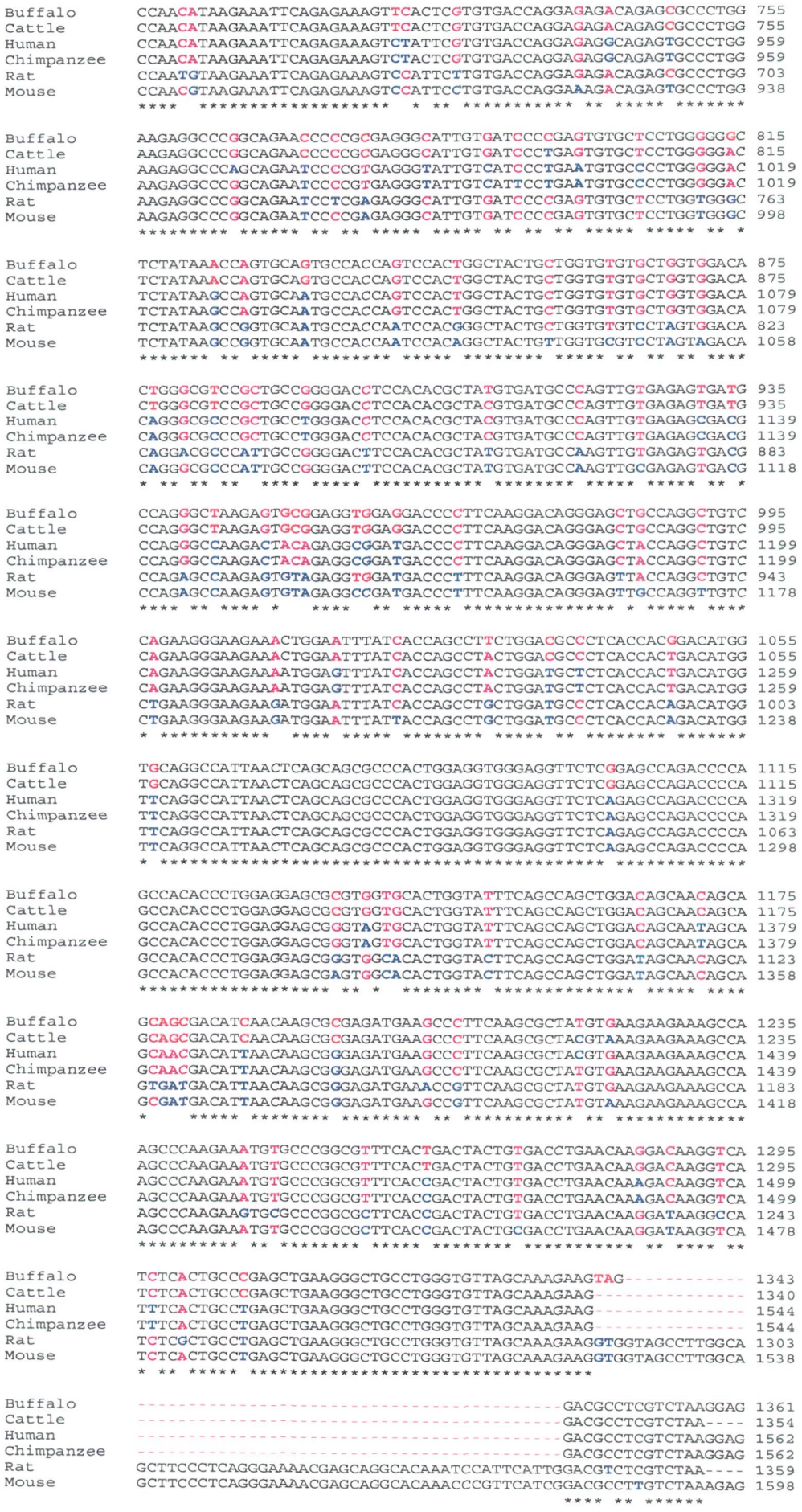


Figure 59. Multiple nucleotide sequence alignment of *Smoc*-1 from different mammals. Some alterations were specific to buffalo or cattle (red) and many were either similar to human/chimpanzee (Pink) or to mouse/rat (Blue). Note more than 90% nucleotide sequence conservation across the mammalian species.

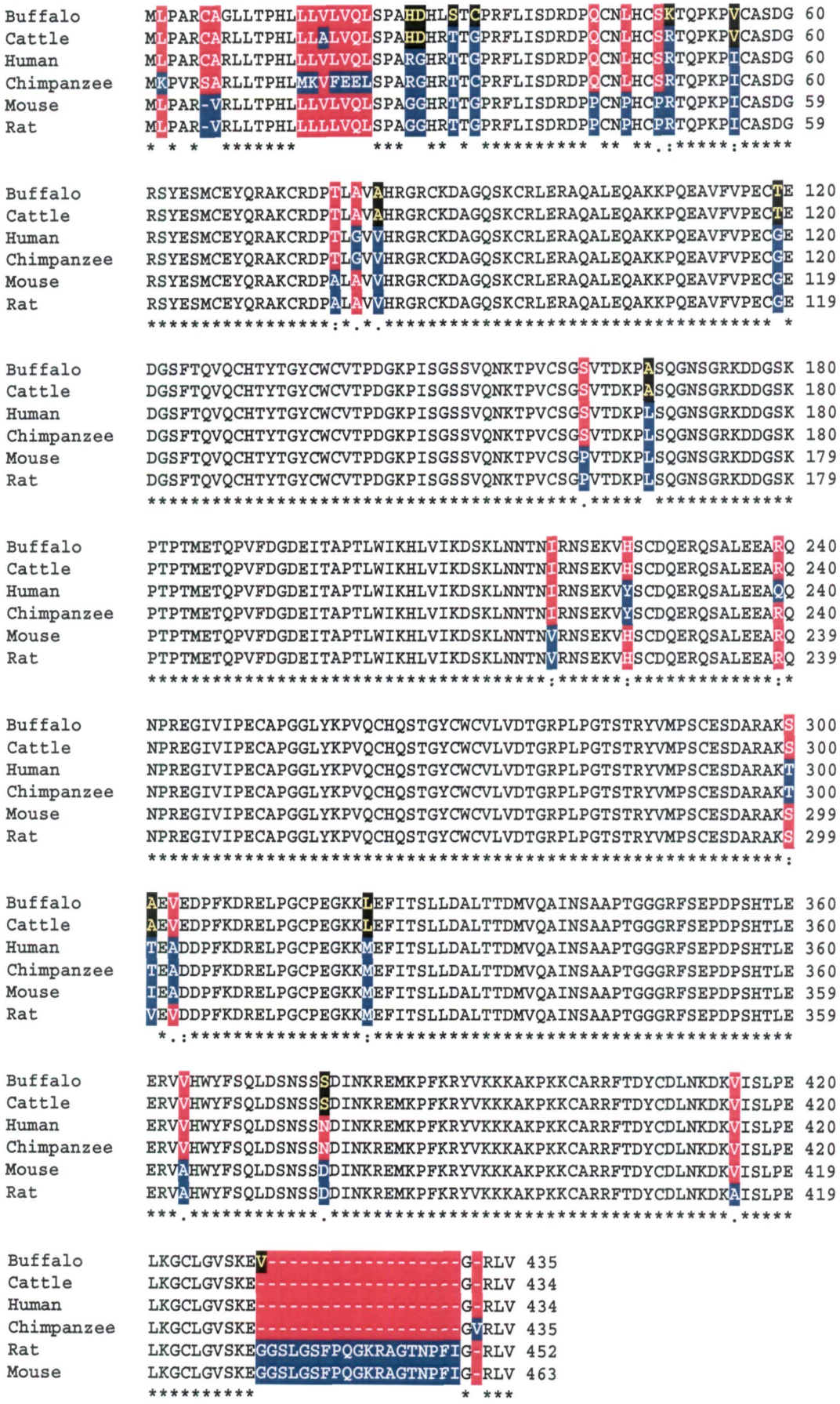
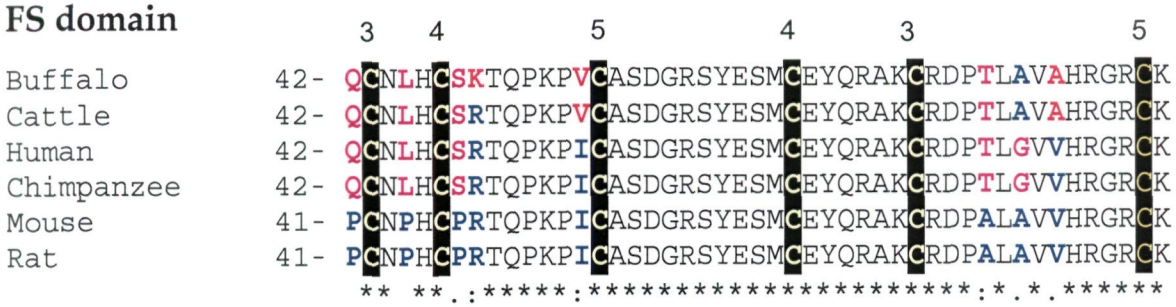
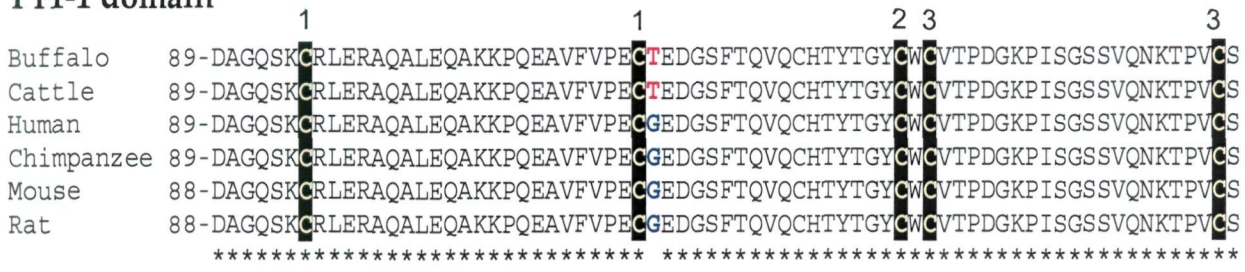


Figure 60. Multiple amino acid sequence alignment of Smoc-1 from different mammals. Some alterations were specific to buffalo or cattle (red overshadowed) and many were either similar to human/chimpanzee (black) or to mouse/rat (Blue). Note more than 85% sequence conservation across the mammalian species.

(A) FS domain



(B) TY1-1 domain



(C) Smoc-1 domain

(D) TY1-2 domain

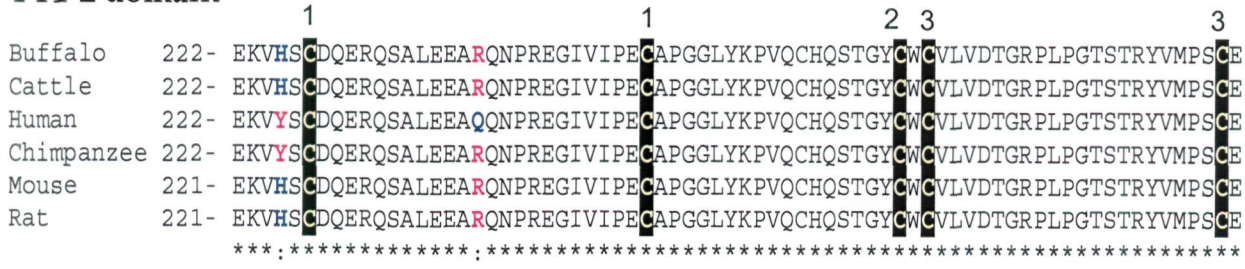


Figure 61

Contd/-

(E) EC domain

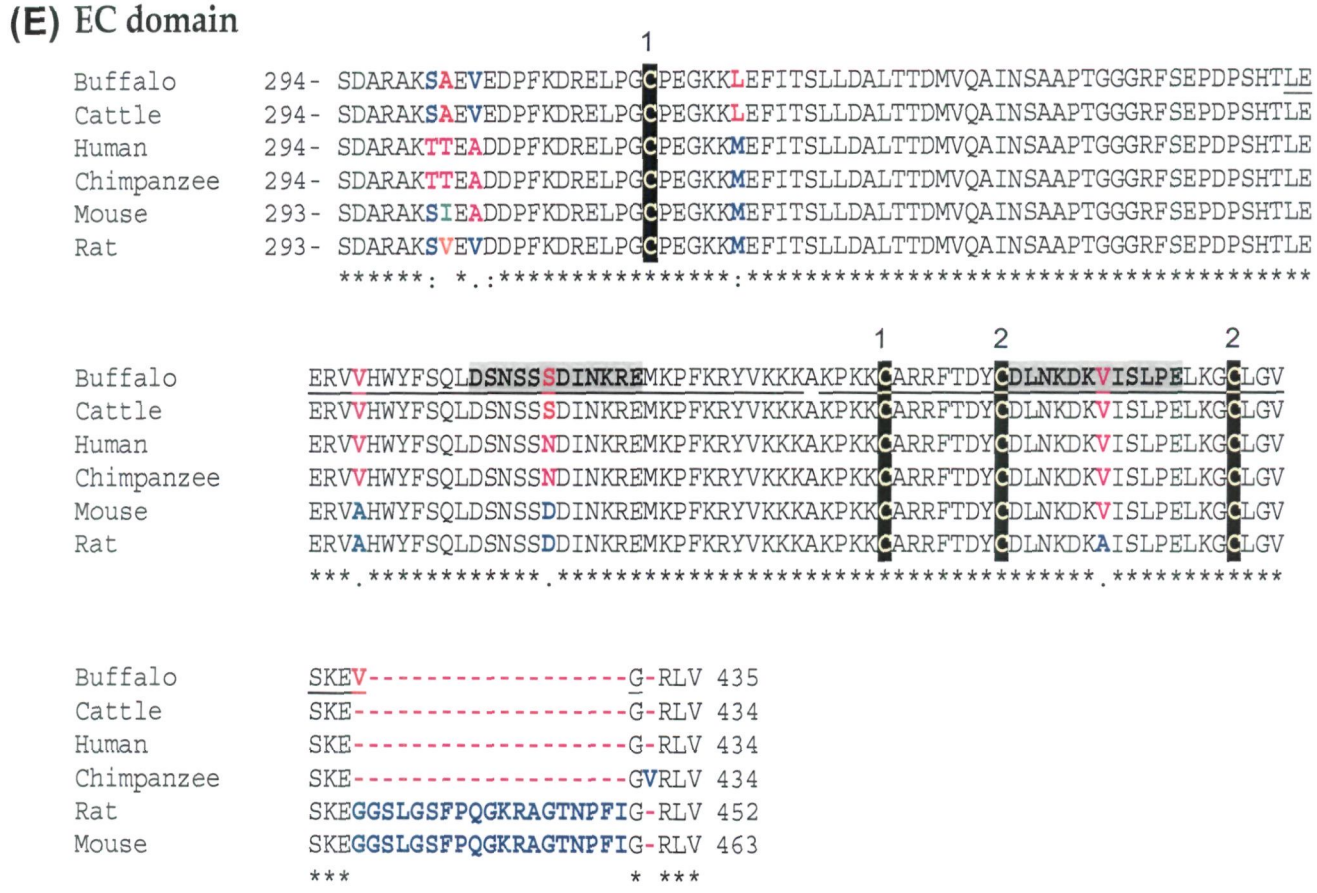


Figure 61. Structure-based alignment of the Smoc-1 protein from different mammalian species. The sequences were aligned across the species for FS **(A)**, TY1-1 **(B)**, Smoc-1 **(C)**, TY1-2 **(D)** and EC **(E)** domains. Mutational hotspots in buffalo *Smoc*-1 are shown red boldface. Most of the observed changes in buffalo *Smoc*-1 were shared either by cattle or human. The "→" indicates the potential *N*-glycosylation site. Pairs of numbers above the sequence correspond to cysteines indicating the predicted disulphide bonds based on the disulphide linkage of BM-40 and thyroglobulin. Both EF hand motifs are underlined and calcium coordinating residues are overshadowed grey. Conserved amino acids are indicated by the stars below and cysteines with a black background.

(A)

Feature	Output summary
SignalP 3.0	Most likely cleavage site between pos. 24 and 25: SPA-HD
ProP 1.0	No popeptide cleavage sites predicted
TargetP 1.1	No high confidence targeting predition
NetPhos 2.0	34 putative phosphorylation sites at positions 37 58 62 65 93 159 161 168 172 179 221 226 233 291 294 300 351 356 373 375 377 417 428 30 119 140 163 276 358 405 63 69 287 407
NetOGlyc 3.1	5 putative O-glycosylated sites at positions 155 182 184 187 345
NetNGlyc 1.0	1 putative N-glycosylated site at position 214
TMHMM 2.0	No TM helices predicted

(B)

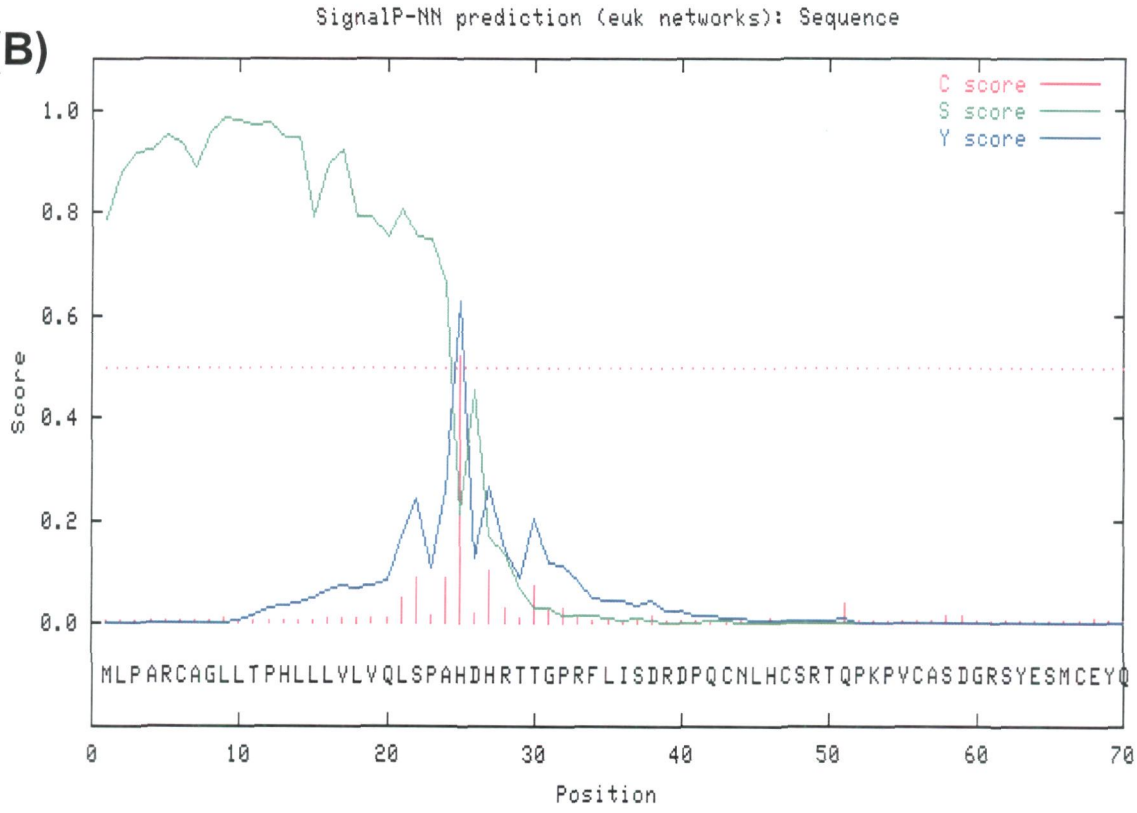


Figure 62. *In-silico* prediction for the presence of signal peptide, N-glycosylation site and O-glycosylaion sites in Smoc-1. Table showing the number of glycosylation and phosphorylation sites **(A)** and the picture showing the signal peptide consensus **(B)**.

the *in-silico* analyses, in contrast to the ECD of BM-40, both EF-hand motifs of Smoc-1 were found canonical indicating the calcium binding sites (Figure 61E).

Compared to other species, buffalo Smoc-1 showed some specific alteration like R49K in FS domain and V431 insertion in ECD. Some of these changes were from polar to non-polar amino acids and vice versa, similar to that in mouse/rat. In case of human/chimpanzee, these changes always maintained similar biochemical nature (Figure 61). Interesting enough, the predicted secondary structure(s) of buffalo Smoc-1 showed alterations at the N-terminus involving replacement of 8 alpha-helices by equal number of beta-sheets and insertion of 3 helices in FS domain compared to that in other mammals (Figure 63). Moreover, few alterations at the N-terminal residues were also observed in the predicted tertiary structures of the Smoc-1 in buffalo compared to that in other mammals (Figure 64).

4.2.2.5 *Single copy of the Smoc-1 gene located on chromosome 11 in the buffalo*

The copy number of the *Smoc*-1 gene was calculated by absolute quantitation assay using SYBR green chemistry in Real Time PCR. A straight curve was drawn using 10 fold dilution series of the known number of plasmid molecules and buffalo genomic DNA. Ct increase of 3.3 per dilution and a single dissociation peak indicated maximum efficiency and high specificity of the primer sets.

Extrapolation of this standard curve demonstrated the single copy status of the *Smoc*-1 per haploid genome in buffalo (Figure 65). Chromosomal mapping of the same was also performed using Fluorescent in *situ* hybridization (FISH) which demonstrated the *Smoc*-1 gene to be present on the distal end of the acrocentric chromosome 11 in buffalo (Figure 66).

4.2.2.6 *Recombinant expression of Smoc-1*

Affinity-purified recombinant Smoc-1 expressed in *E. coli BL21 (DE3)* revealed a major band at ~ 70 kDa in 10% SDS-PAGE under

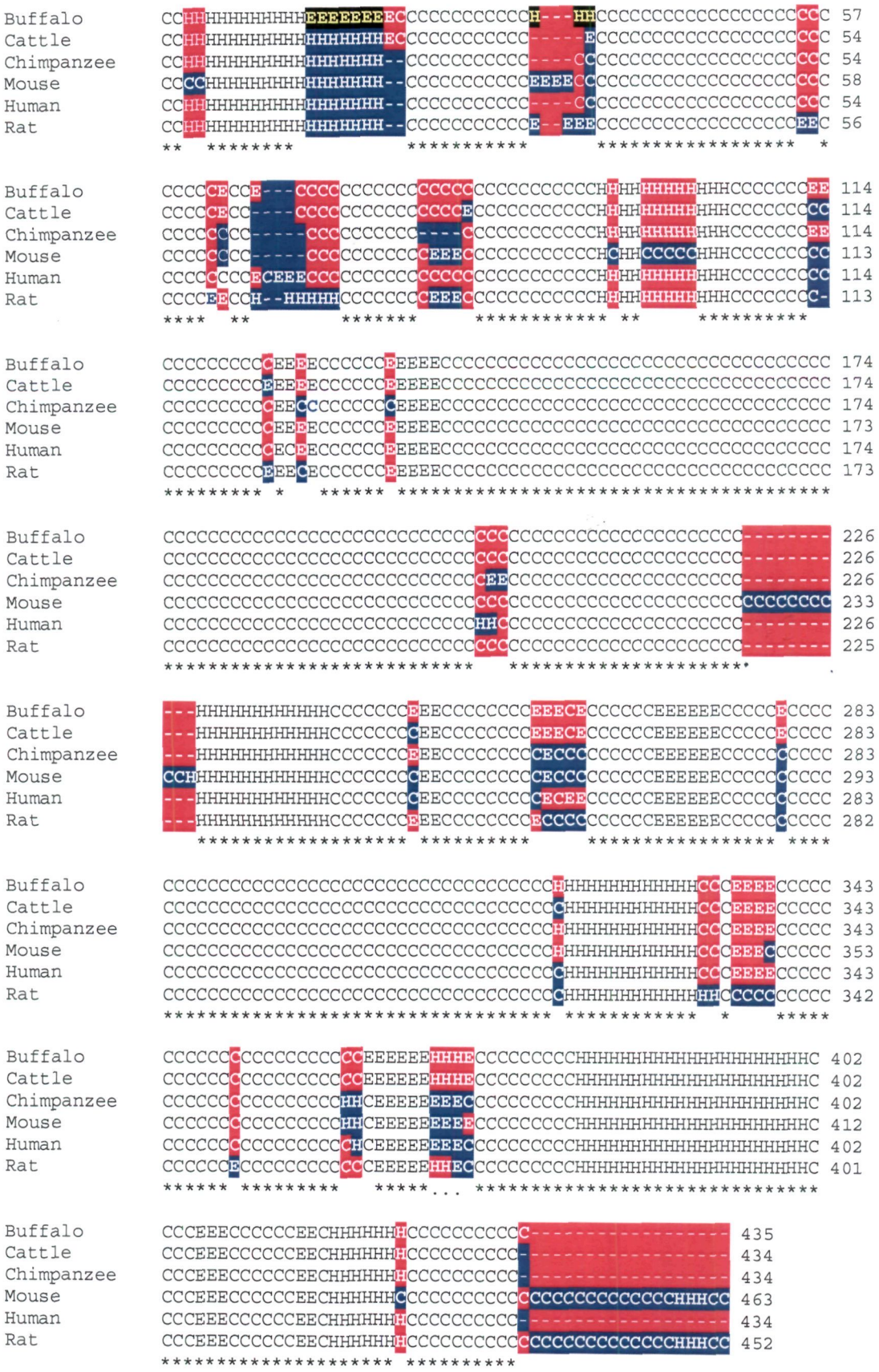


Figure 63. Predicted secondary structure(s) of Smoc-1 protein from different species. The replacement of helix formed by 8 residues with the beta-sheets, insertion of helices and the minor alterations throughout the protein are shown in black. Changes similar to cattle or human and the ones similar to rat/mouse or chimpanzee are shown in red and blue backgrounds, respectively. The additional coils and helices in mouse and rat are due to bigger coding frame of Smoc-1 in these species.

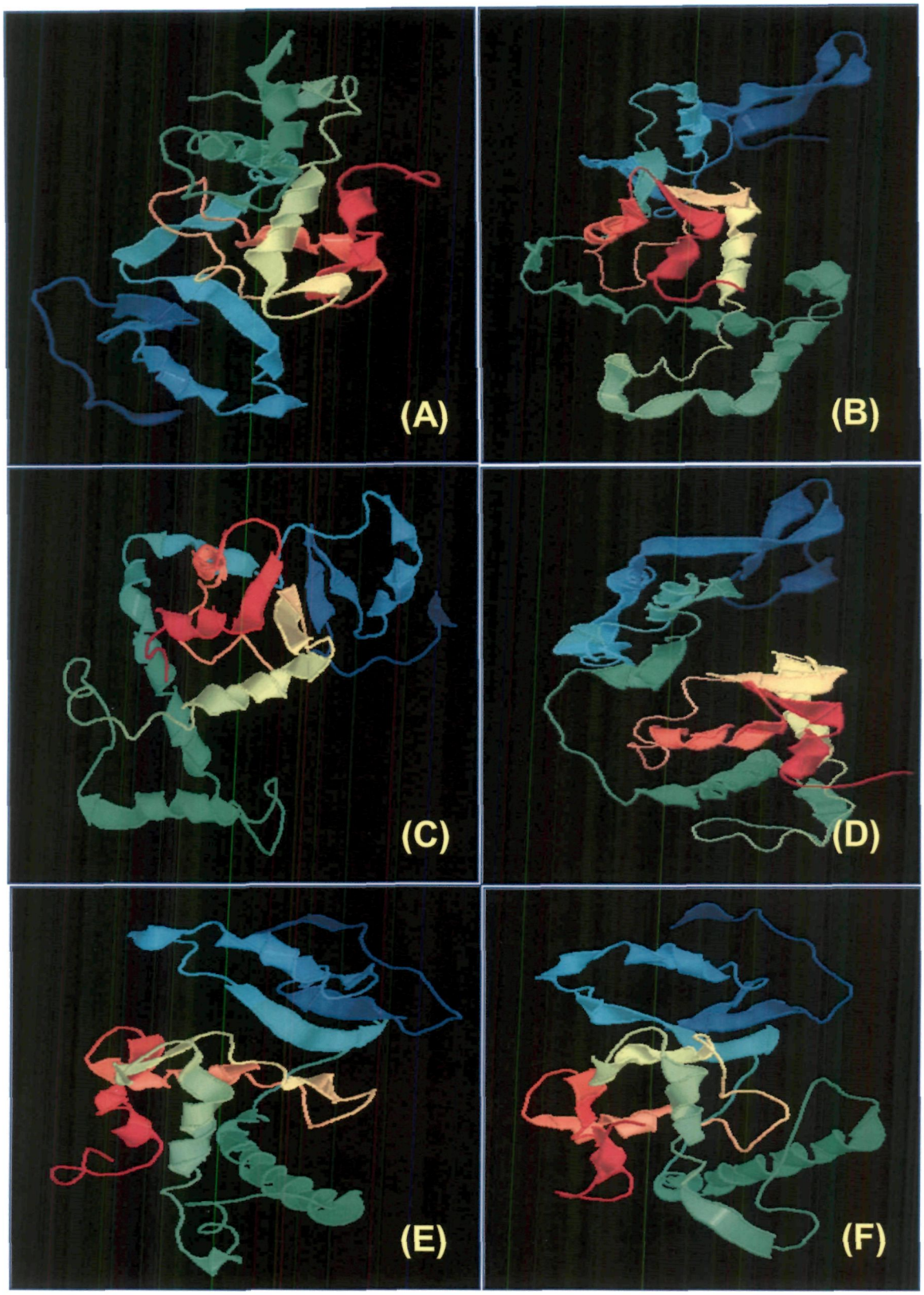


Figure 64. Predicted tertiary structures of Smoc-1 protein from different species; Buffalo **(A)**, Cattle **(B)**, Human **(C)**, Chimpanzee **(D)**, Mouse **(E)** and Rat **(F)**. Please note the alteration at N-terminal of Smoc-1 in buffalo and human in comparison to cattle and chimpanzee.

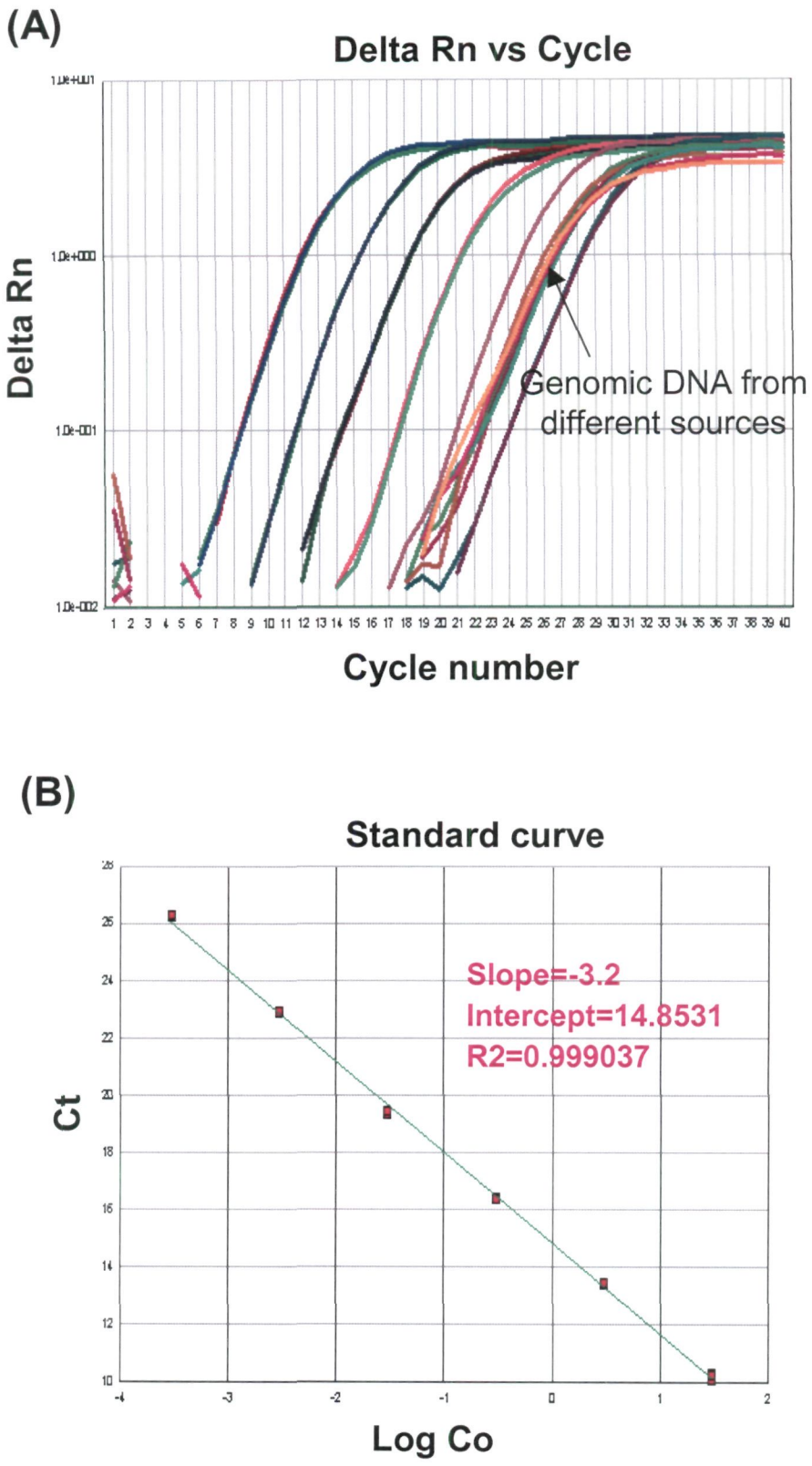


Figure 65. Copy number calculation of *Smoc*-1 gene. Real Time PCR amplification plot based on ten fold dilution series of F*Smoc*-1 recombinant plasmid **(A).** Genomic DNA from blood of male/female buffalo and semen samples used as template (A) to obtain a standard curve using SYBR Green assay **(B)** which detected the single copy status of this gene. The value of R2, slope and intercept are given in the standard curve.

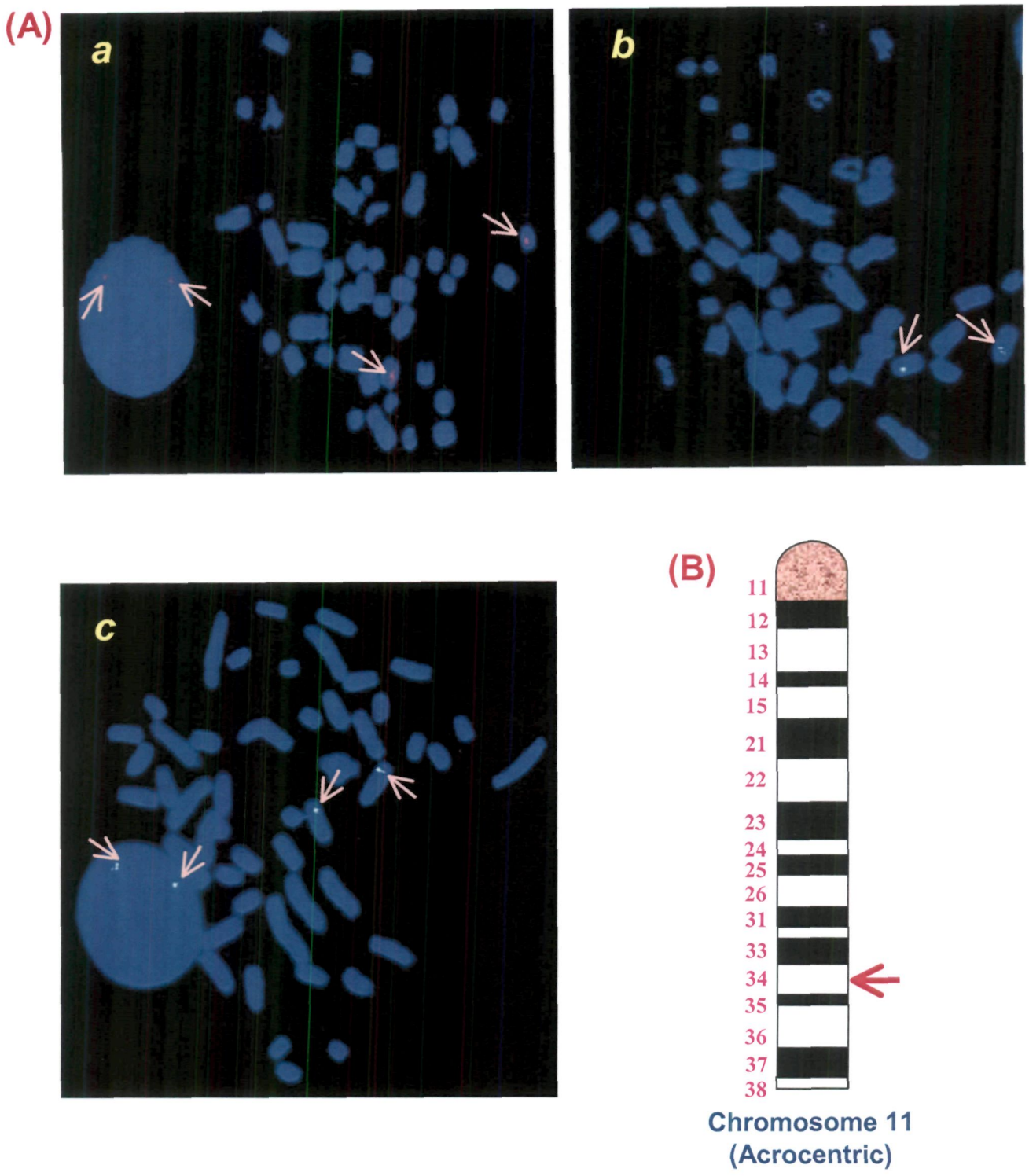


Figure 66. Chromosomal mapping of *Smoc*-1 gene at chromosome 11 in buffalo. Fluorescence *in situ* hybridization demonstrating the presence of *Smoc*-1 gene on the distal arm of the acrocentric chromosome 11 **(A, panels a-c)** and detailed mapping of this gene with respect to its position on the G-banded ideogram following the chromosome nomenclature standardized by ISCNDB, 2000 **(B)**.

reduced conditions (Figure 67). The deduced molecular mass of the mature Smoc-1 is 45.4 kDa and the remaining ~25 kDa represented GST tag. The Anti-PSmoc1-pAb recognized native protein of ~ 70 kDa in the western-blot analysis. Anti-SySmoc-1-pAb generated against commercially synthesized 26 amino acids (69S to 95G) also showed the similar results (Figure 68), substantiating the high specificity of these antibodies. The pre-immune serum did not detect any protein in the Western blot analysis (not shown).

4.2.2.7 *Highest expression of Smoc-1 transcript variants in liver*

Northern blot analyses showed abundant *Smoc*-1 transcripts in liver and faint signals in the testis and ovary. After prolonged exposure, the spleen, lung, kidney and heart also showed negligible to faint signals (Figure 56). Similarly, RT-PCR followed by Southern hybridizations detected reduced signals in spleen (Figure 69A-B) and negligible ones in lung, kidney and heart after prolonged exposure (not shown). Using quantitative expressional analysis, β-actin as an internal control and lung cDNA as calibrator, the highest level of expression of transcript variant-01 (165-364 folds) and -02 (360-697 folds) was observed in liver (Figure 69C-D). This was substantiated further by expression data from the five additional animals. In the same assay, sperm cDNA showed similar transcription level of *Smoc*-1 as that of testis. However, relatively higher amount (1.2-3.5 folds) of variant-02 was observed in all the tissues examined as compared to that of variant-01 (Figure 69E). Based on these observations, the variant-02 may be addressed as "major transcript" and -01 as "minor" one.

Since all the mRNA transcripts may not translate into protein, the relative quantitation mentioned above was substantiated by the Western blot analysis using total tissue proteins and anti-SySmoc1-pAb which detected the ~45 kDa bands (mature protein). Total protein was isolated from different tissues and checked on 10% SDS PAGE for their quality (Figure 70). Thereafter, equal amount of protein for each tissue was taken for the western blot analyses. Interestingly, the strongest signal was detected in liver and faint ones in testis, ovary and spleen (Figure 69F)

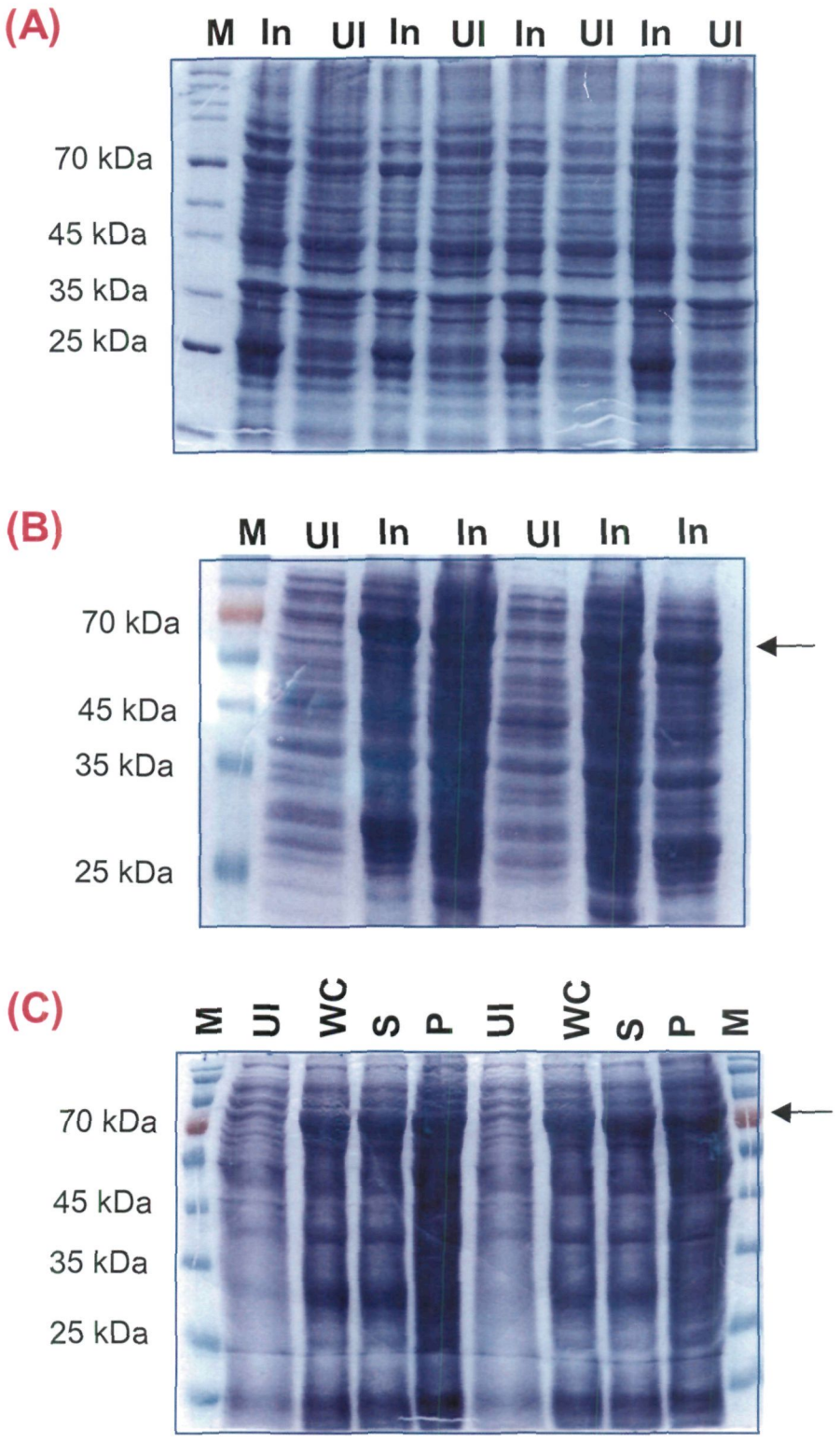


Figure 67. SDS-PAGE showing the un-induced and induced Smoc-1 protein **(A).** Please note the difference in the IPTG induced bands of 70 KDa representing the for recombinant pGEX-4T1-PSmoc1 containing Smoc-1gene and the GST protein of 25 KDa. This induction was checked *multiple times and at multiple temperatures* for e.g. 25-37 deg C **(B).** The recombinant protein was then sonicated and the crude protein, supernatant and pellet was loaded at the gel to check the protein folding **(C).** M denotes the marker, UI for uninduced, I for induced, WC for total crude protein, S for supernatant and P for pellet. Equal amount of protein was detected in both soup and pellet.

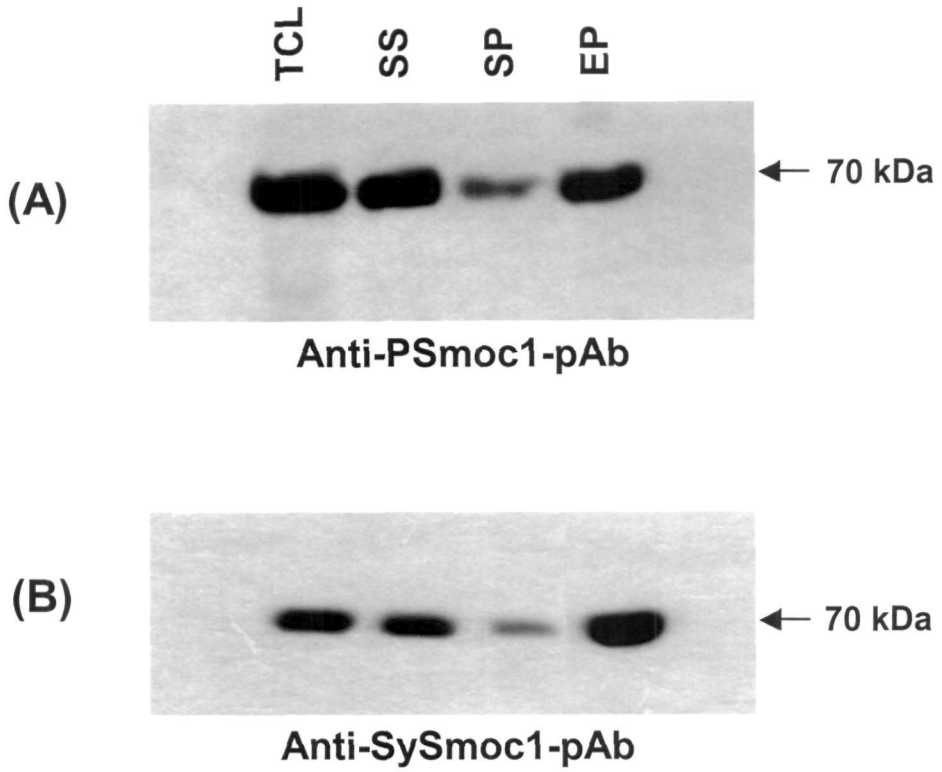
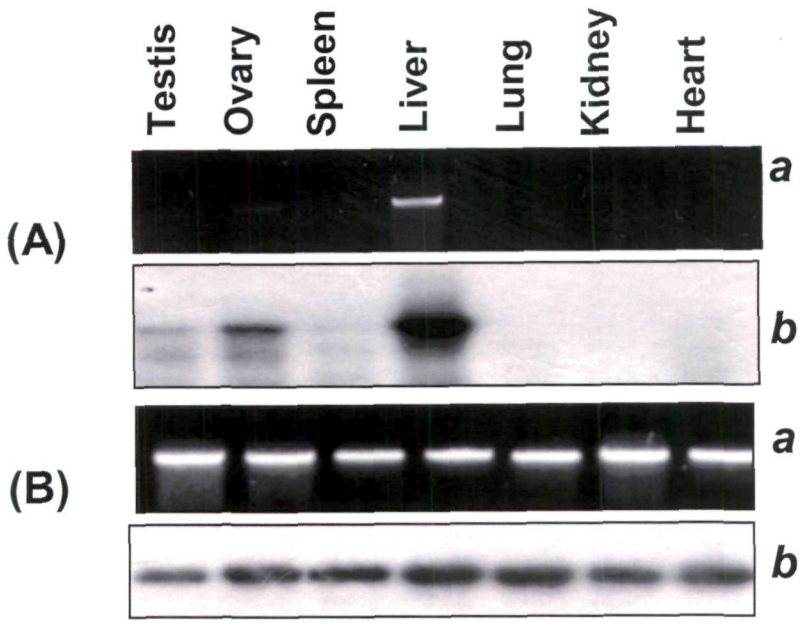
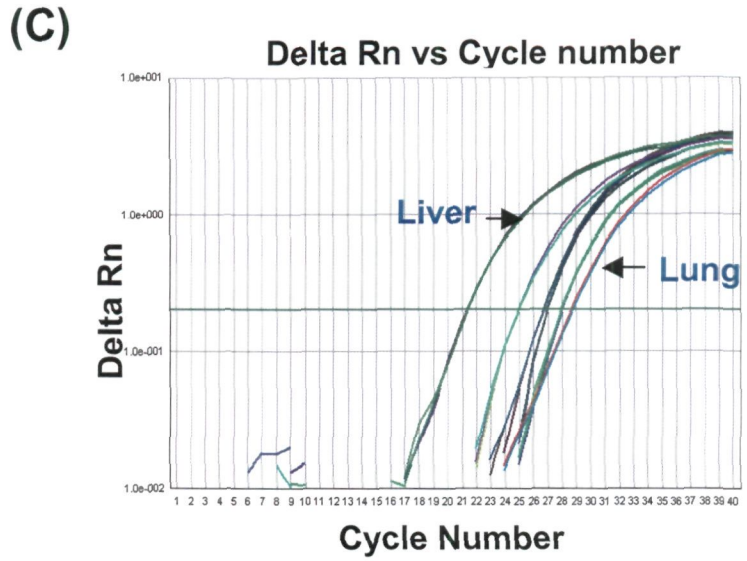


Figure 68. Western blot using anti-Smoc-1 antibodies. Anti-P*Smoc*1-pAb specifically generated against GST-*Smoc*1 recombinant protein showed ~70 kDa protein in western blotting **(A)**. The same results were observed using Anti-Sy*Smoc*-1-pAb generated against the synthesized amino acids specific to *Smoc*-1 unique domain **(B)**. TCL denotes total cell lysate; SS, sonicated supernatant; SP, sonicated pellet and EP, eluted protein.





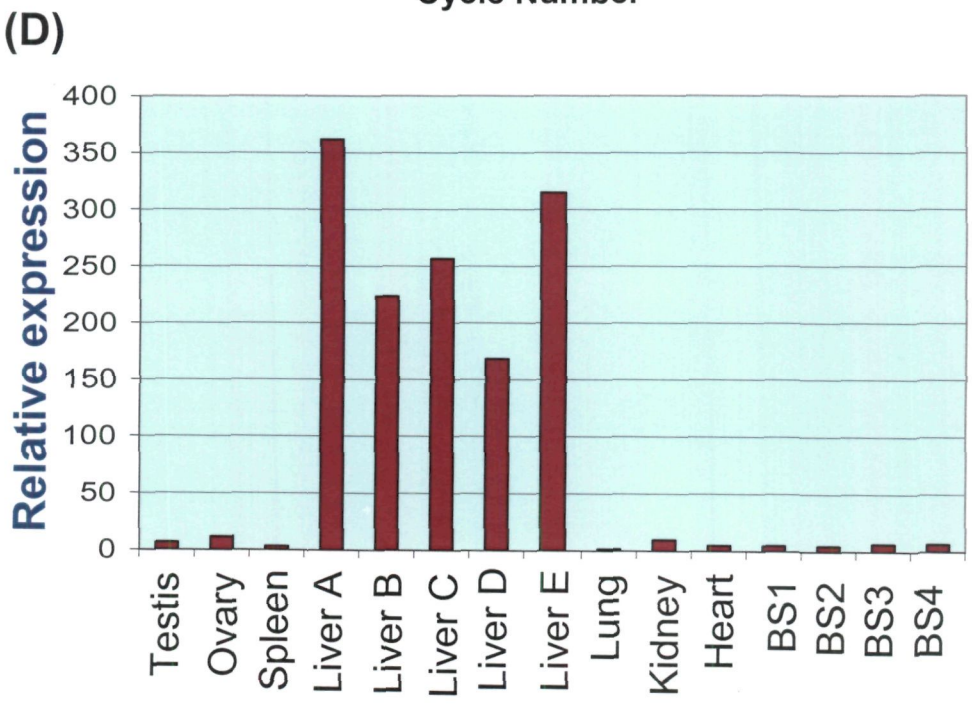


Figure 69

Contd/-

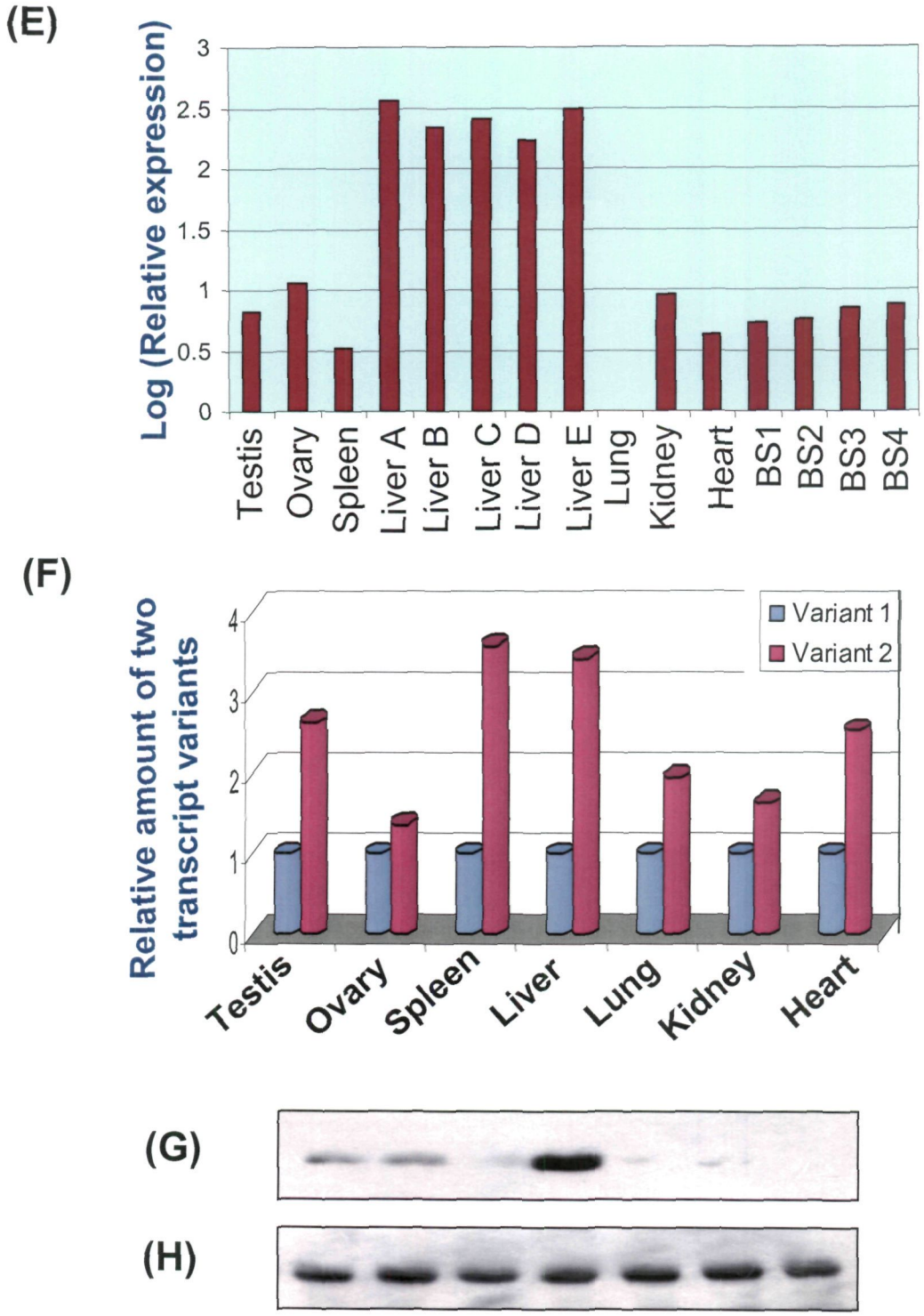


Figure 69. Highest expression of *Smoc*-1 in liver. RT-PCR showed exclusive band only in liver but Southern hybridization detected reduced signals in testis, ovary and spleen **(A)** whereas control *β*-actin showed almost equal intensity signal in each tissue **(B).** Quantitative expression of *Smoc*-1 by Real Time PCR confirmed maximum expression (163-364 folds) in liver in five different animals compared to that in lung **(C-E)**. Note the buffalo spermatozoa from four different animals (BS1-4) also showed 4-7 folds transcripts in different animals, similar to that in testis. Relative quantitation also demonstrated higher expression (1.2-3.5 times) of variant-02 compared to that of -01 in each of the tissues examined **(F)**. Western blotting with anti-Sy*Smoc*1-pAb **(G)** and anti-*β*-actin-mAb as positive control **(H)** confirmed the highest expression of Smoc-1 protein in liver.

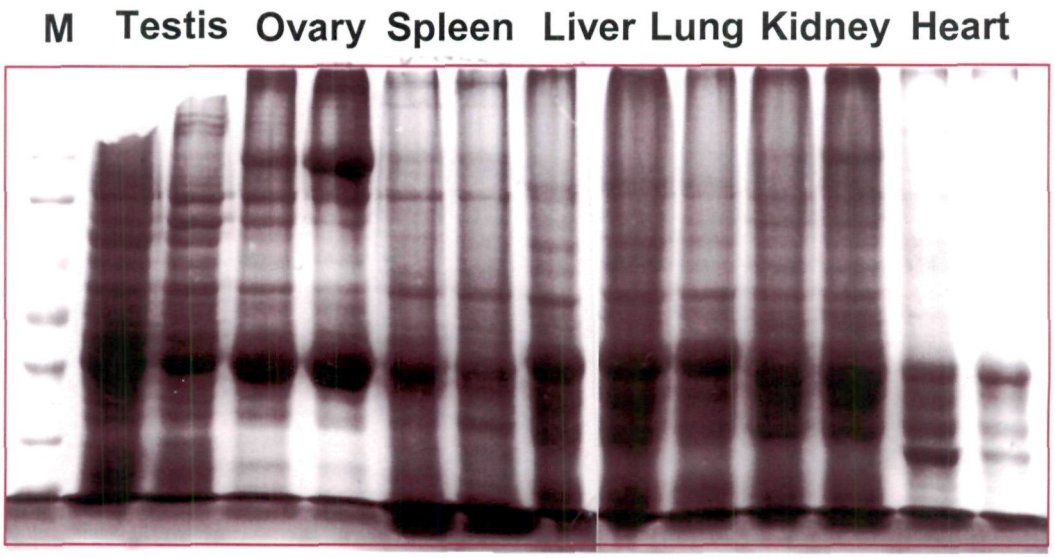


Figure 70. SDS-PAGE showing total protein isolated from different tissues of buffalo.

whereas lung, kidney and heart were found to be devoid of Smoc-1 protein corroborating the RT-PCR and relative expressional analysis data. As a control, anti-β-actin-mAb showed almost equal signal intensity in each of the tissues studied (Figure 69G).

4.2.2.8 *Age specific expression profile of the Smoc-1 in water buffalo*

We report for the first time, the expression profile of *Smoc*-1 in water buffaloes of varying ages starting from 20 days to 15 years. Lowest expression of *Smoc*-1 was detected in the blood lymphocytes of animals aged 20 days with gradual increase in the expression (1.5-2 times) from 1 month-10 months. However, a sharp enhancement in the expression (25-30 times) was detected at the age of 1-1.25 years and this remained consistent up to the age of 15 years and beyond (Figure 71A). The dramatic increase in expression of *Smoc*-1 at around 1 year of age was further confirmed by western blotting using anti-SySmoc1-pAb and total protein isolated from blood samples of the same animals (not shown).

The comparative expression analysis of the two transcript variants revealed a gradual increase of the variant-02 compared to that of variant-01 with the progression of age (Figure 71B). The expression of variant-01 was higher in animal aged up to ~6 months, after which the variant-02 expression starts increasing gradually during the ages of 6 to 15 months and remained consistent thereafter.

4.2.2.9 *Association of Smoc-1 with the basement membrane*

The immunohistochemical studies were performed in order to localize the Smoc-1 onto different tissues. Smoc-1 was found to be present abundantly in the basement membrane zone of discontinuous endothelial cell layer or the tunica media around the central veins in liver (Figure 72A-D). In addition, Smoc-1 was also found to be ubiquitously distributed in the connective tissues surrounding each lobule and extracellular matrices. In testis, Smoc-1 was abundant in the basement membrane zone surrounding coiled seminiferous tubules below the columnar epididymis and scarcely in the interstitial tissues (Figure 73A-D).

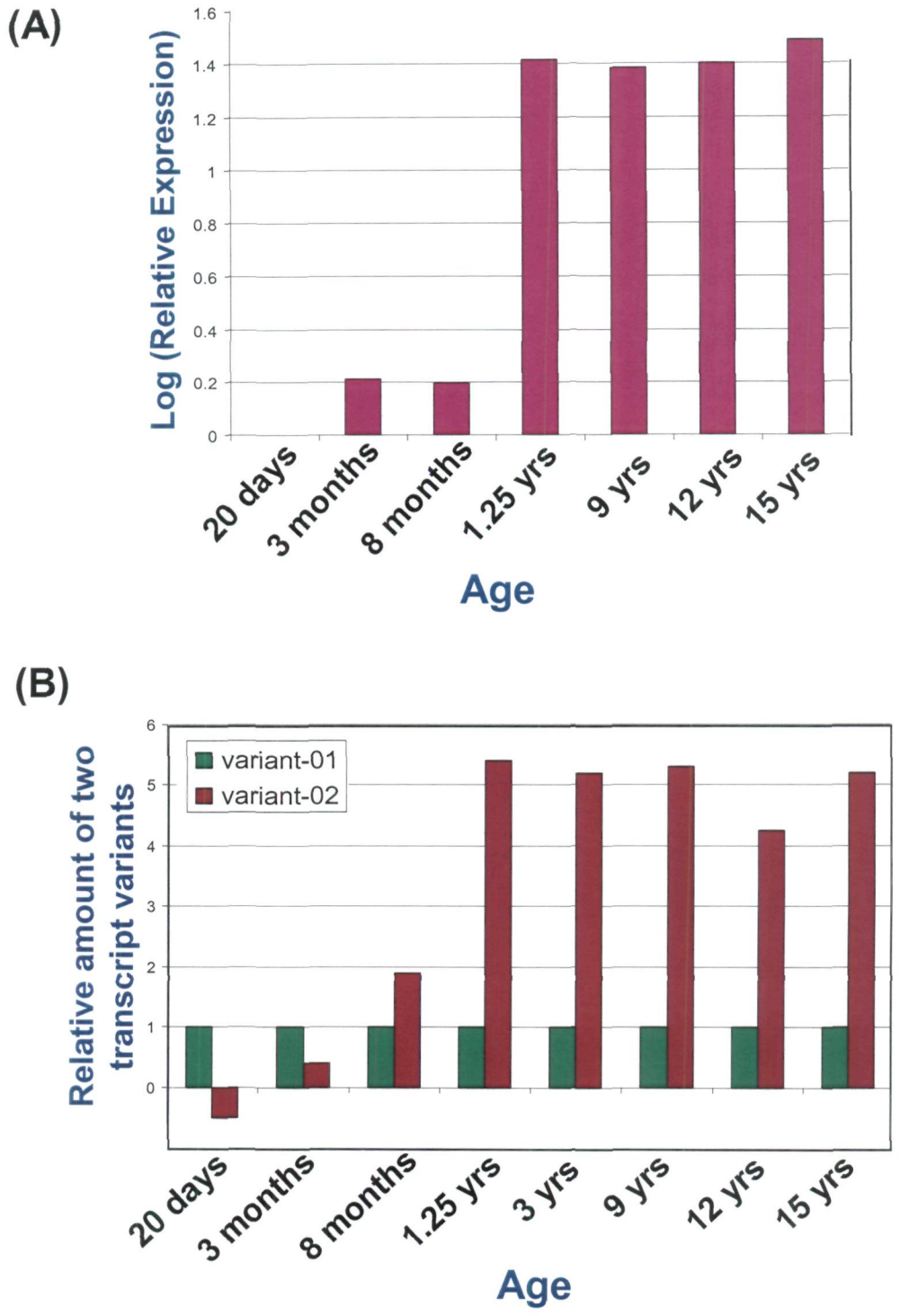


Figure 71. The quantitative expression carried out using cDNA samples isolated from blood lymphocytes of different age group of animals has been shown in (G). Note the markedly increased expression of Smoc-1 in animals 10 months and beyond. Relative quantitation showing higher expression of variant-01 in the animal up to ~6 months of age and that of variant-02 in the animals of age 6 months and beyond (H).

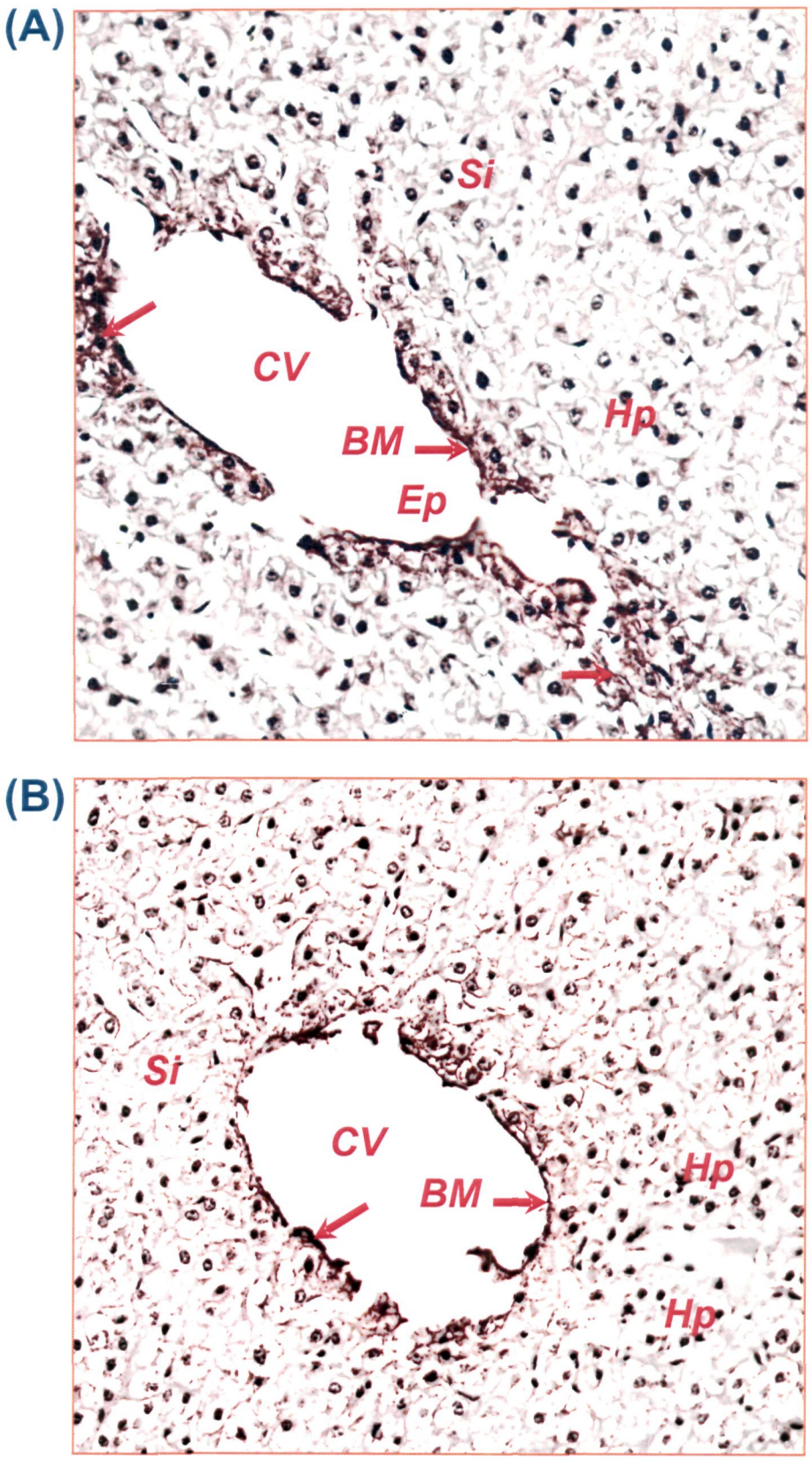


Figure 72

Contd/-

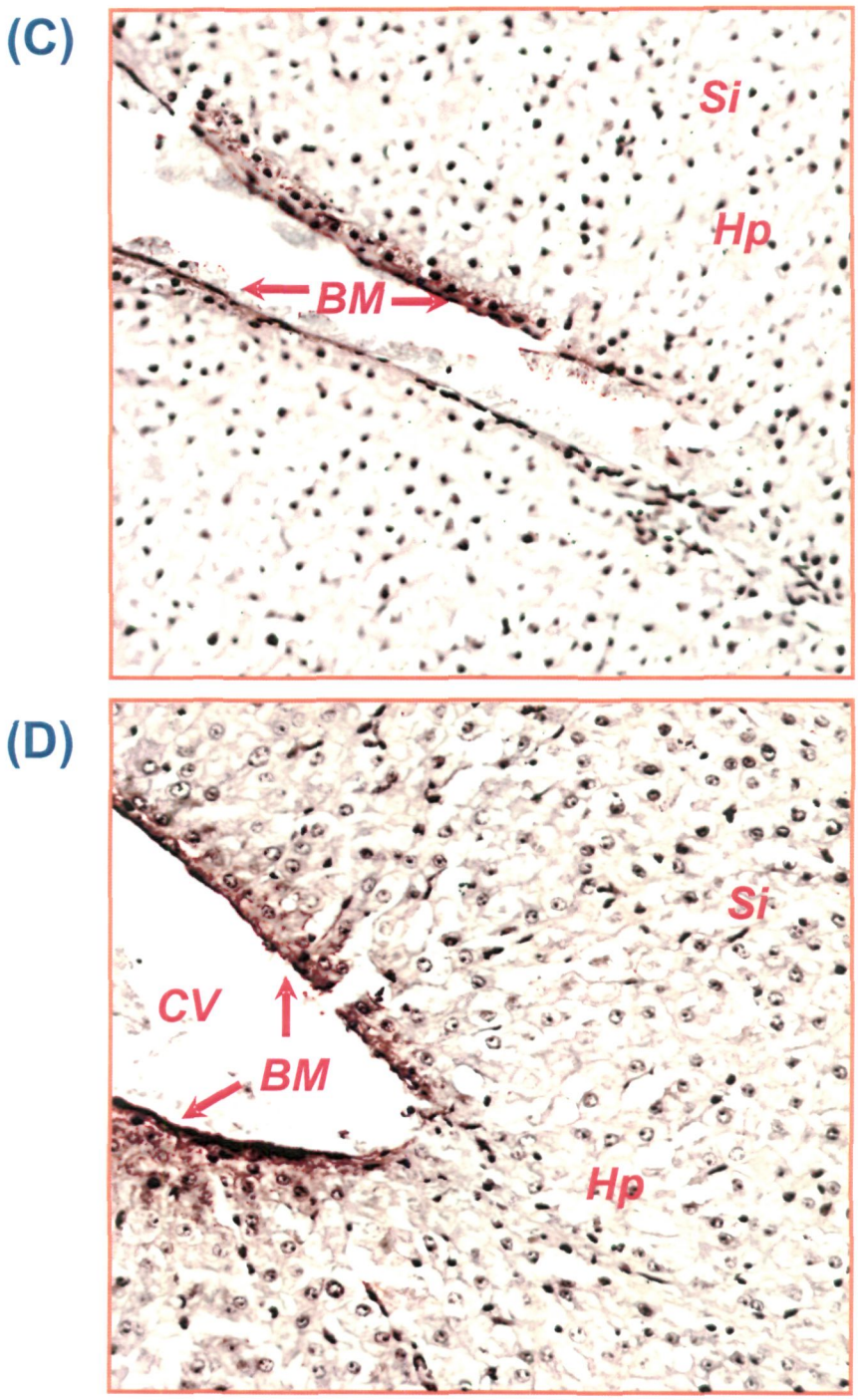


Figure 72. Indirect Immunohistochemistry of buffalo tissue sections using Anti-SySmoc-1-pAb. Distribution of Smoc-1 in basement membrane zone of endothelial cell layer and extracellular matrices of liver **(A-D).** Note the localization of Smoc-1 protein in the basement membrane zones indicated by red arrows. The bars represent 5 µm in panels **A-D**. *Hp* denotes hepatocytes; *Si*, sinusoids; *CV*, central vein and *Em*, discontinuous endothelial cell of central vein in 'A', and; *BM*, basement membrane zone.

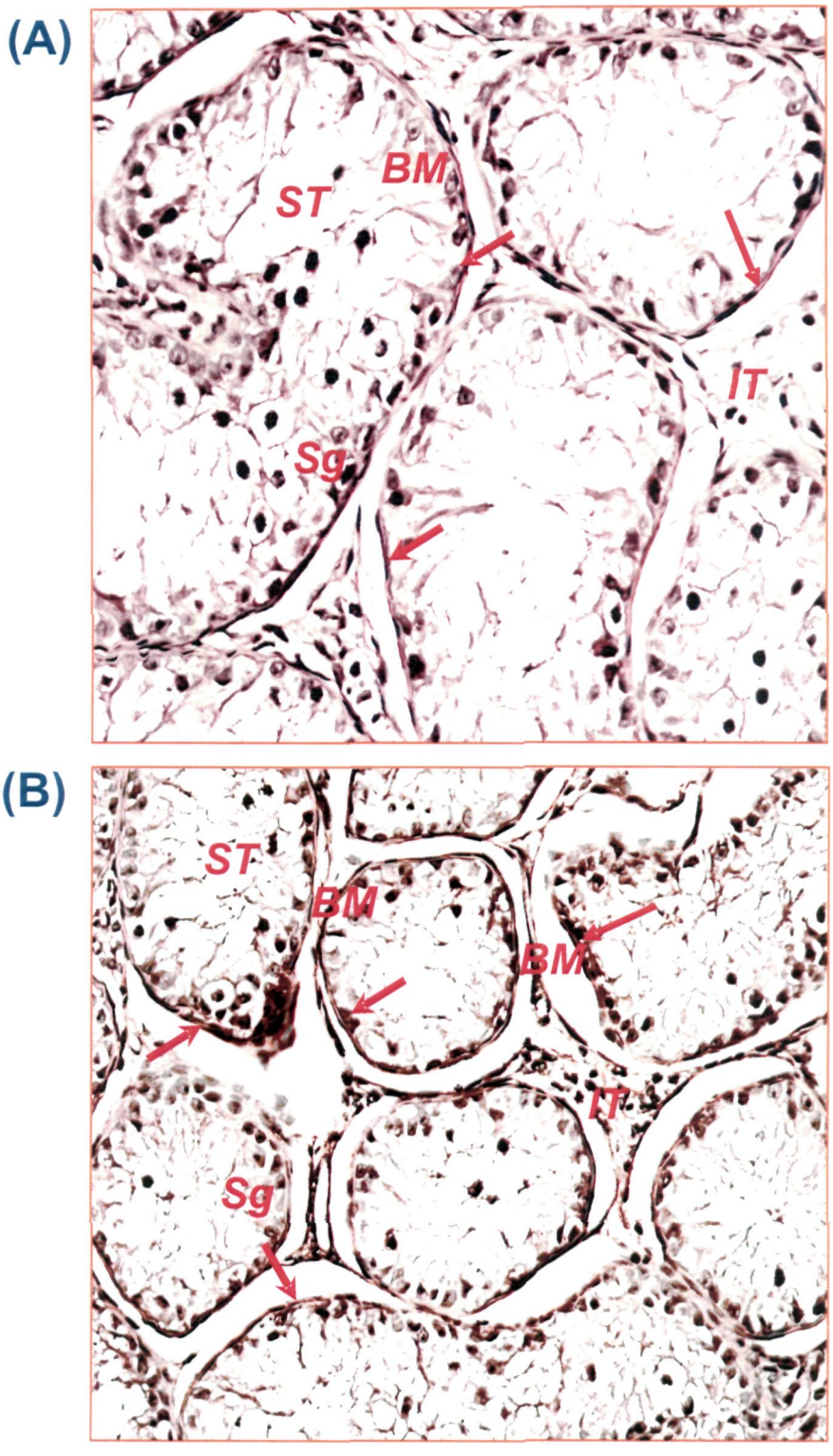


Figure 73

Contd/-

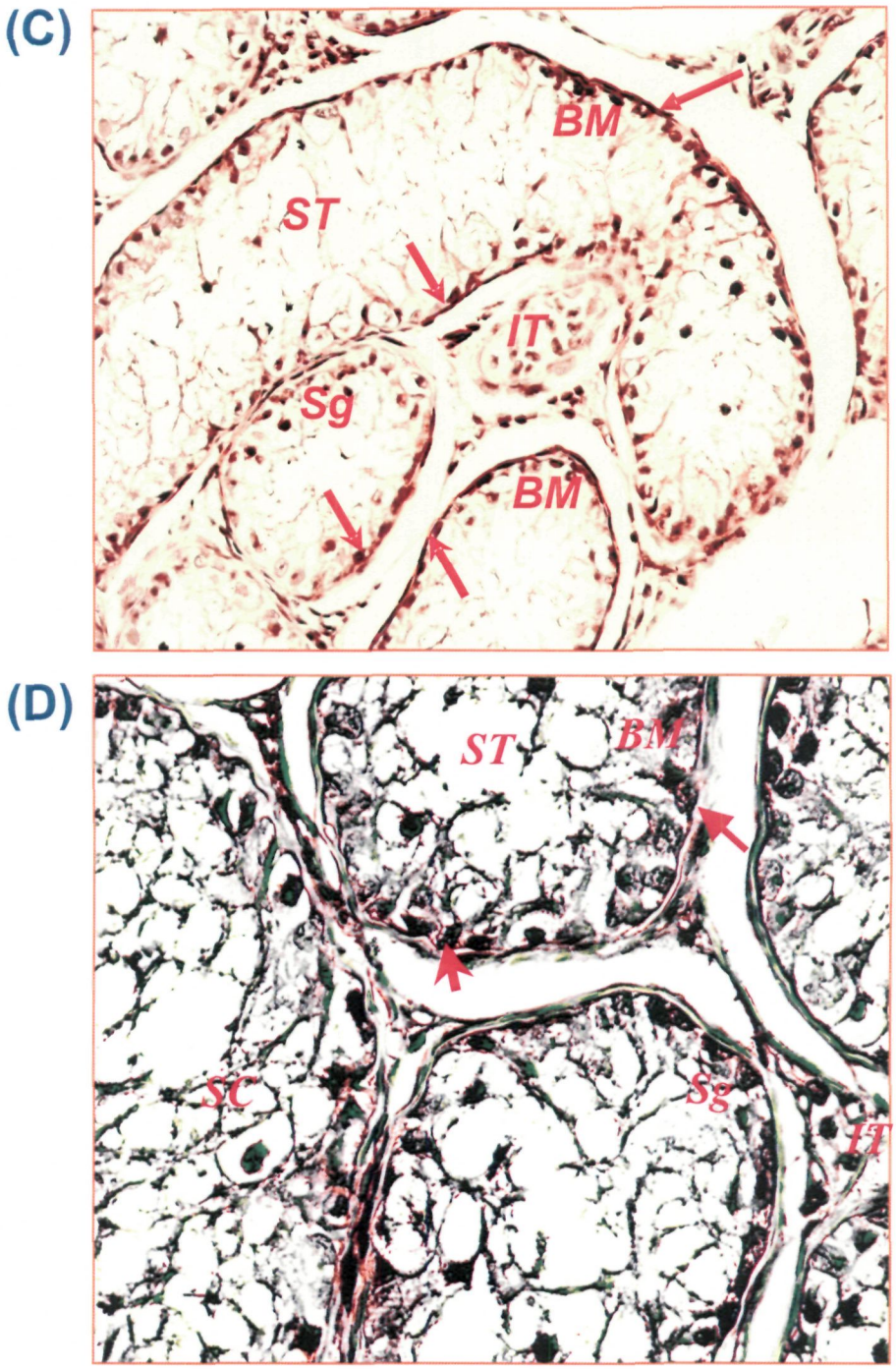


Figure 73. Indirect Immunohistochemistry of buffalo tissue sections using Anti-SySmoc-1-pAb. Distribution of Smoc-1 in basement membrane zone of endothelial cell layer and extracellular matrices of liver **(A-D)**. Note the localization of Smoc-1 protein in the basement membrane zones indicated by red arrows. The bars represent 5 µm in panels **A-D**. *Hp* denotes hepatocytes; *Si*, sinusoids; *CV*, central vein and *Em*, discontinuous endothelial cell of central vein in 'A', and *ST,* seminiferous tubules; *BM*, basement membrane zone; *Sg*, Spermatogonia and *IT*, interstitial tissues .

Smoc-1 was localized in the zona pellucida of ovary in buffalo (not shown) similar to that in mice. In other tissues also, it remained localized within the basement membrane zones. In conclusion, the Smoc-1 was construed to be a basement protein similar to that in other species.

DISCUSSION

5. DISCUSSION

Simple sequence repeats (SSRs) are abundant in the non-coding regions (Toth *et al.*, 2000; Katti *et al.*, 2001) and have been implicated with the gene regulation, chromosomal fragile sites and genome imprinting (Katti *et al.*, 2001) but the biological significance of their association with the coding genomes remains largely unresolved. However, few reports had suggested the vital roles of these repeats, residing/hidden in the mRNA transcripts, in gene regulation (Cummings and Zoghbi, 2000; Li *et al.*, 2004), but the unresolved distribution and organizational/expressional variations of tagged transcript diversity remains limitations of these studies. In the present study, we established the association of the GACA, GATA and 33.15 repeats with the somatic as well as spermatozoal transcriptomes of water buffalo, *Bubalus bubalis*. In addition, the observed sequence polymorphisms and differential gene expression suggested the diverse functions of these repeat-tagged transcripts in different cell types, ages, stages and tissues. Moreover, highest expression of the GACA/GATA tagged transcripts in testis and/or spermatozoa indicates their crucial roles in male gametogenesis. Below we also discuss the distinct recruitment of GACA/GATA repeats along with the distribution of 33.15 repeat loci in different species.

The potential biological significance of the candidate genes *Smoc*-1 and *c-kit,* which were targeted for their detailed characterization to gain an insight into its structural and functional organization, and expressional status in water buffalo, *Bubalus bubalis* is also discussed. Both of the genes showed tissue-, age- and species-specific organization and expression.

5.1 Repeat tagged transcript diversity

5.1.1 Global distribution of the 33.15, GACA and GATA repeats

Satellite sequences, initially thought to be of no biological consequences have been found to be transcribing in a number of species

(Kizawa *et al.*, 2005). Our study showed the distribution of the 33.15 sequences in the flanking regions and within the mRNA transcripts of many structural, functional and regulatory genes in buffalo (Table 10). In addition, a detailed database search (www.ncbi.nlm.nih.gov/blast) demonstrated that six other species were found to have a minimum of 14 to 16 nucleotides of 33.15 repeat loci in their mRNA transcripts (Table 9). Of these, cattle showed 2, human 13, rat 4, mouse 8, dog 6 and pig 1. It is likely that many more mRNA transcripts, though still not part of the GenBank, may harbor these sequences.

Extensive *in silico* analysis demonstrating total absence of GACA/GATA repeats in prokaryotes and their presence in different eukaryotes studied suggested their accruement in the non-coding and coding genomes of eukaryotes (Table 6-8). Further detailed analysis of *S. cerevisiae*, *C. elegans*, *Arabidopsis thaliana* and *Drosophila melanogaster* revealed that their genomes harbored only few or no GACA/GATA repeat which substantiated the gradual accumulation of these repeats in higher eukaryotes with the course of evolution. Exploration of the GACA/GATA tagged transcriptomes from lower to higher eukaryotes showed absence of GACA repeats in *Arabidopsis thaliana, Dictyostelium discoideum*, *Drosophila melanogaster* and *C. elegans,* and GATA in *Sus scrofa*, *C. elegans* and *D. discoideum.* Moreover, their presence in respective non-coding genomes suggested their species-specific distribution and distinct functions to maintain genetic integrity. However, GACA/GATA was found to be present in transcriptomes of most of the higher eukaryotes (Table 7-8). These repeats seem to have been acquired in the transcriptomes with the increase in the genetic complexities in higher eukaryotes. Most of these transcripts tagged with the GACA and GATA repeats had been characterized as the functional ones. Further, the chromosome-wise distributional studies for these repeats (Figure 4) highlighted their concentration on the sex chromosomes of different species. Thus, the sex-chromosomal occurrence and diversity of tagged transcripts suggested the involution of the GACA/GATA repeats in functional regulation of the sex determination.

5.1.2 Biological significance of the repeats within the somatic and spermatozoal transcriptomes

Tandem repeats residing within the coding regions are mostly involved in transcription/translation, can also mediate phase variation, and alter the functions and antigenecity of the proteins encoded (Vergnaud and Denoeud, 2000; Jordan *et al.*, 2003). In this study, a total of 63 different mRNA transcripts (34, GACA-tagged; 10, GATA-tagged; and 19, 33.15-tagged) representing few known and most of the novel ones were identified (Figures 7-9; Table 10-12). Thus, present study can be used as a basis for using other repeats in order to demonstrate the combined significance of such repeats within and adjacent to the coding regions. The above discussed species-specific distribution of these repeats became more interesting when these repeats picked up various mRNA transcripts in the buffalo spermatozoa as well. Spermatozoa have come a long way from being regarded as an artifact to its repository of a variety of RNAs including siRNAs. The initial controversy which shrouded their existence has now been cleared by the identification of several transcripts in the spermatozoa of different species including human (Miller *et al.*, 2005; Krawetz *et al.*, 2005).

Although, many signaling molecules and transcription factors have been reported to pass by the spermatozoan into the zygotic cytoplasm on fertilization (Krawetz *et al.*, 2005; Miller *et al.*, 2005; Ostermeier *et al.*, 2005), yet ~3000-5000 transcripts remains to be characterized in the spermatozoa. The existence of the SSRs tagged transcripts in the buffalo spermatozoa is the first finding which suggests the involvements of the 33.15, GACA and GATA repeats and their tagged mRNA transcripts during the pre- and post-fertilization events. It became more significant, when all these mRNA transcripts, uncovered by the 33.15, GACA and GATA repeats, showed faithful evolutionary conservation across thirteen different species (Figure 27-29), suggesting broader significance of these repeats in these species and may be in all the eukaryotic species.

5.1.3 Implications of the organizational variations in repeat tagged transcripts

Of the 7 transcripts uncovered with the consensus of 33.15 repeat loci, the 846 bp one showed random nucleotide changes in the somatic tissues that did not alter the amino acids. However, gonads (ovary and testis) showed changes at six identical places resulting in conspicuous alterations and deletions of the amino acids in the C-terminal region (Figure 16). Biological significance of such alterations and deletions are not clear nor is the underlying mechanism. Perhaps, this could be a case of programmed sequence modulation which is restored back some times after fertilization since this was detected in other animals as well. The 846 bp fragment showed homology with Adenylate Kinase Like gene which is known to play an important role in reproduction (Luconi *et al.*, 2001). How the so-called programmed sequence modulation is involved with the event(s) of reproduction is not clear nor do we know the stage(s) at which such changes take place. Nucleotide changes in 602 bp fragment at two places in heart seem to reflect real mutations and not polymorphism since the same was found in other animals also (Figure 17).

The buffalo transcriptome was found to be enriched with GACA repeat while other species including human were GATA rich. The outcome can be explained by fact that 34 mRNA transcripts were isolated with the GACA repeat and 10 by GATA repeat (Figure 8-9; Table 11-12). Although the primates and cetartiodactyls' genomes are relatively GC poor (Duret *et al.*, 2002), the GC richness of buffalo genome and transcriptome seem to be unique for its organization and thus for replication timings, genetic recombination, methylation and gene expression. The differential transcript profiles uncovered here by the GACA/GATA repeats may be explained either towards their various functions in somatic tissues, gonads (testis/ovary), and spermatozoa, or differential functions at various stages of development. Absence of GATA-tagged transcripts in lung and heart is anticipated to be transcriptional quiescence of the representative genes (Figure 9). Several other tissue-specific transcripts uncovered by GACA/GATA entail their exclusive requirement in those respective tissues. Moreover, there are two possible explanations for the detection of 20 of 34

GACA-tagged and 6 of 10 GATA-tagged transcripts in testis/spermatozoa. First, the transcripts could not be picked up in other tissues due to either polymorphic nature of the STRs or lower number of transcripts, and alternatively, they are dormant in other tissues barring testis and spermatozoa. The homology search establishing the novel status of ~40% GACA-tagged and all the GATA-tagged transcripts (Table 11-12) further corroborated the species-specific distribution of these repeats and are thus picked up the unidentified or uncharacterized genes in buffalo. However, involvement of these repeat tagged genes either in signal transduction or cell-cell interaction pathways indicated their crucial roles in various cellular functions essential for the life cycle of a cell.

DNA sequence variation can contribute to phenotypic variation by affecting the steady-level of mRNA molecules of a particular gene in a given cell or tissue (Kliebenstein *et al.*, 2006). The tissue- and spermatozoa-specific sequence organizations in ~30% of the GACA/GATA-tagged transcripts (Figures 19-26) substantiated this hypothesis. Few tissue-originated transcripts such as the GACA-tagged 1.8 Kb and 1.3 kb, and GATA-tagged 800 bp transcripts harbored major nucleotide variations suggesting their distinct organization and significance in the spermatozoa, lung and spleen, respectively. As suggested by the Real Time PCR values, the sequence insertion in the 1.8 kb transcript seems to cease or reduce its expression in the lung, whereas the expression level of the 1.3 kb transcript was enhanced by nucleotide insertion in the spermatozoa (Table 14). However, the expression level of 800 bp transcript was not affected by the sequence alterations in spleen. Some transcripts showed nucleotide changes exclusively in the spermatozoa, few in testis, whereas in several others, variations were shared only between testis and spermatozoa for instance, the ANKD26 (Figure 23). These findings may be explicated by the silenced state of the representative genes in the somatic tissues which are active in testis and spermatozoa or vice versa. Several other nucleotide variations to a particular tissue suggest their different putative involvements in those respective tissues.

5.1.4 Prospects of repeat tagged transcripts carrying highest expression in the testis and spermatozoa

Sequence polymorphisms have been shown to regulate the differences in gene expressions, and inter- or intra-specific phenotypic variations in various organisms (Carrol, 2000; Duret *et al.*, 2002). Therefore, the quantitative expressional studies carried out for the uncovered repeat tagged transcripts explored the positive significant expressional variation in all the somatic/gonadal tissues and spermatozoa. Of the seven 33.15 tagged transcripts, four (AKL, LRRN6A, Spergen-3 and TCRGL) showed highest expression in testis (Figure 33). The expression profiles so observed suggested the potential implications of these transcripts in various testicular functions. This is an important observation because following this approach, genes expressing preferentially in gonad(s) may be easily accessed. *Smoc*-1 and TCRL genes showed highest expression in liver and spleen respectively (Figure 33). *Smoc*-1 is calcium binding unique matricellular glycoprotein expressed by different cell types and associated with development, remodeling, cell turn-over and tissues repair mechanism (Brekken and Sage, 2000; Vannahme *et al.*, 2002). The 576 bp transcript showing partial homology to TCRL-α gene seems to have immunological significance as demonstrated by its highest expression in spleen. The LRRN6A gene encoding for a transmembrane leucine-rich repeat protein involved in axonal guidance, migration, nervous system development and regeneration processes of neuronal cells (Carim-Todd *et al.*, 2003) also showed maximum expression in the testis.

The uniform expression of the about 30% GACA-tagged transcripts (Figure 34) suggested their consistent requirements in all the tissues and spermatozoa, whereas ~10% with highest expression in the liver or spleen indicated their putative involvement in the hepatocellular and immunological activities, respectively. Interestingly, the highest expression of a total of 29 transcripts comprising 19 GACA- and all the 10 GATA-tagged transcripts in testis and/or spermatozoa (Figure 34-35) corroborated their deep involutions in the spermatogenesis and fertilization events. The negligible or lower expression of these gene fragments in

other tissues including ovary, further substantiated their potentials in the male gonad development.

Of all the spermatozoa specific transcripts, highest expression was identified for ankyrin repeat domain-26 which mediates the cell-cell interaction and histone modification pathways (Mosavi *et al.*, 2004; Wang *et al.*, 2005). The WASF2 gene which is involved in changes in cell shape, motility or function (Takenawa and Suetsugu, 2007), also showed highest expression in spermatozoa proposing its involution in both spermiogenesis and fertilization events. In addition, few other transcripts such as Ubap1 also showed testis specific expression. Though ubiquitin associated domain is thought to be involved in ubiquitinization pathways (Qian *et al.*, 2001), the testis specific expression of Ubap1 observed here suggested its pivotal role in various testicular functions. Few studies have hypothesized the participation of GACA/GATA repeats in reproduction and heterogametic germ cell development (Gangadharan *et al.*, 2001; Singh *et al.*, 1994). Present study demonstrating the testis- and spermatozoa-specific expression of majority of the GACA/GATA tagged transcripts further substantiates this hypothesis.

5.1.5 MASA and comparative genomics

Our detailed study indicates that the GACA, GATA and 33.15 repeat sequences are present within several mRNA transcripts involved in several pathways such as signal transduction, transcription, translation, immunological activities, and sex-differentiation, besides the non-coding regions. MASA mediated approach seems to be highly effective for isolating a large number of mRNA transcripts which harbor the consensus of these repeats. However, a number of repeats can be used for MASA to establish their combined conclusive significance within and adjacent to the coding regions. The functional studies of the transcripts so uncovered will resolve the enigma of such simple sequence repeats in the mammalian genome. Thus, in the context of comparative genomics, mRNA transcripts commonly expressing in a large number of species may be segregated. Following this approach, genes with highest levels of expression in a given tissue may be easily identified and information so obtained from different

breeds of buffalo may be collated to characterize its unknown genome which will help finally to establish the genetic basis of eliteness or other physical and physiological attributes of this animal.

5.2 Differential organization and expression of *Smoc*-1 and *C-kit* in buffalo

5.2.1 *C-kit* and its tissue specific nature

5.2.1.1 *C-kit: structure and domain organization*

The pleiotropic proto-oncogene *c-kit* receptor belongs to transmembrane receptor tyrosine kinases (RTK) family type-3 (Chabot *et al.*, 1998; Andre *et al.*, 1992) similar to the receptors for platelet derived growth factor (PDGF) and macrophage-colony-stimulating factor (M-CSF) (Qiu *et al.*, 1988). The buffalo c-kit glycoprotein also includes an immunoglobulin like extracellular, a single transmembrane and an intracellular tyrosine-kinase domain similar to that in human (Galli *et al.*, 1992; Gokkel *et al.*, 1994). *C-kit* gene has been reported to be single copy in all the species studied thus far. In order to ensure that multiple forms of mRNA transcripts detected in buffalo are not originating from the pseudogenes and that this is indeed a single copy, we used Real Time PCR to assess its copy number status. Single copy status of *c-kit* gene in buffalo correlates with that in other mammals.

It is well studied that the exons and exon-intron boundary regions of this gene are conserved but extracellular domain and intronic sequences show variations across the species (Reith *et al.*, 1991; Crosier *et al.*, 1993). Similarly, several unique features were observed in the buffalo *c-kit* receptor compared to that in other mammals (Figure 42-43, 47). Moreover, full length c-kit peptide was detected in testis whereas other tissues showed its truncated version, possibly due to nucleotide changes at several places (Figure 44-45). The truncated c-kit protein devoid of ECD and TMD did not contain ATP binding site and is responsible for generating the soluble form. This makes a complex with the ligand in the extracellular matrix which does not allow signaling across the nucleus

(Besmer *et al.*, 1986; Besmer, 1995). The truncated c-kit protein is known to play crucial roles in post-meiotic haploid cells during spermiogenesis. The coexistence of full length and truncated tyrosine kinases in buffalo testis is a remarkable characteristic feature corroborating with the earlier results (Kierszenbaum, 2006). However, the truncated peptide lacking Intracellular and transmembrane domains detected in other tissues was never reported earlier. Presence of truncated peptide in the ovary and all the other tissues except testis was indeed startling. Tissue specific nucleotide changes in the mRNA transcripts perhaps reflected their programmed sequence modulation, a possible mechanism of transcriptional inactivation.

5.2.1.2 *Tissue specific alternate splicing and c-kit*

Despite organizational variations within the extracellular domain, the *c-kit* gene undergoes tissue and stage specific alternative splicing (Serve *et al.*, 1995) and almost all the species studied thus far have been reported to show alternate splicing resulting in its variant forms (Thommes *et al.*, 1999). These variant forms of mRNA transcripts are different from the mutant ones. However, negligible information is available on the mutant mRNA transcripts related to a specific function in any species barring rats (Kierszenbaum *et al.*, 2006; Prasanth *et al.*, 2004). Presence of a 618 bp mRNA transcripts in heart, ovary and testis highlights its pressing requirement in these tissues (Figure 46). The same transcript representing part of the tyrosine kinase domain was detected in post-meiotic haploid cells in human and mice testis, and human mature spermatozoa (Paronetto *et al.*, 2004) indicating a conserved role of this protein in gamete function. The mRNA transcripts detected in liver, lung, ovary and other somatic tissues of buffalo are envisaged to be implicated in tissue specific signal transduction.

5.2.1.3 *C-kit with testis specific expression*

Very few reports are available on the mRNA transcripts in spermatozoa (Ostermeier *et al.*, 2005). The highest level of *c-kit* mRNA transcripts in buffalo testis evoked our interest to assess the same in the

semen samples. Discernible expression of *c-kit* in testis and its detection in the spermatozoa (Figure 51) indicated signaling supremacy of this gene in control and regulation of testicular functions leading to normal spermatogenesis. Based on our data, we hypothesize that mutation(s) in testis specific mRNA transcript reported herein may hamper the process of spermiogenesis. This may prove to be a highly reliable marker system for segregating fertile animals from the infertile ones (Prasanth *et al.*, 2004).
A comprehensive study of this gene undertaken in other animals and systematic characterization of individual mRNA transcript would go a long way to uncover the significance of each transcript and its mutant status. This would also resolve if indeed there is a testis specific mRNA transcript involved in regulation of spermatogenesis.

5.2.2 *Smoc*-1 and its transcript variants

5.2.2.1 *Smoc-1 and SPARC family: Comparative organization*

Present study demonstrates the association of the consensus sequence of minisatellite 33.15 with the coding sequence of the *Smoc*-1 which is the member of Basement membrane-40 (BM-40) family. However, the existing significance of this association remained unclear. BM-40 also known as SPARC (Secreted protein acidic and rich in cysteine) is an anti-adhesive secreted matricellular glycoprotein family (Termine *et al.*, 1981; Lane and Sage, 1994) associated with tissue remodeling during normal developmental processes such as angiogenesis and bone mineralization (Brekken and Sage, 2001). This family also includes SC1/Hevin (Guermah *et al.*, 1991; Girard and Springer, 1995), Testican (Alliel *et al.* 1993), tsc36/Flik/FRP (Shibanuma *et al.* 1993) and *SMOC*-2 (Rocnik *et al.* 2006). Till date, *Smoc*-1 has been characterized only in a few mammals showing variations in domain organization. SPARC family proteins are basically characterized by the presence of a follistatin-like (FS) and a C-terminal extracellular (EC) calcium binding domains with two EF-hand binding motifs (Hohenester *et al.*, 1996; Maurer *et al.*, 1995) whereas in *Smoc*-1, the FS and EC domains are separated by the two TY domains which are themselves split by the novel *Smoc*-1 domain (Figure

52 and 61), thus maintaining its organizational uniqueness amongst SPARC family.

5.2.2.2 *Potential implications of multi-domain proteins like Smoc-1*

Predicting the significance of orthologues separated millions of years ago has always been a difficult proposition, especially in the context of multi-domain proteins with frequent insertions or deletions. Thus, biological functions for FS, TY and EC domains in *Smoc*-1 are still speculative. The FS domain is not only the characteristic of BM-40 family, but also found in other follistatin related genes like C6, C7, agrin, and the transmembrane receptors TMEFF1 and TMEFF2 (Eib and Martens, 1996; Horie *et al.*, 2000). Similarly, presence of TY domain in other proteins (Nakada *et al.*, 2001) makes it difficult to ascertain its function in the *Smoc*-1. The TY contains six cysteine residues including a characteristic CWCV tertrapeptide (Molina *et al.*, 1996) which is also conserved in buffalo *Smoc*-1. The high content of aromatic amino acids in the unique Smoc-1 domain suggests the formation of a folded domain with a hydrophobic core. Presence of two EF hand motifs in ECD of buffalo *Smoc*-1 is predicted for its calcium binding affinity as it has been confirmed experimentally using circular dichroism in human SMOC-1 (Gersdorff *et al.*, 2006). Presence of acidic residues at positions 1,3,5,9 and 12, and the helix signatures encompassing the calcium binding loops are also conserved for both EF hand domains in buffalo Smoc-1 (Figure 61).

5.2.2.3 *Smoc-1 and its transcript variants in different species*

Owing to about 90% sequence homology with cattle, human and other species, buffalo *Smoc*-1 showed similar arrangement of various domains. Analysis of the gene structure in buffalo, human and mouse reveals intactness of each domain border (Fitzgerald and Shenk, 1981) maintaining its reading frame even when some exon/intron is inserted or deleted (Figure 60-61). However, a number of specific alterations at nucleotide and amino acid levels were found to be unique to buffalo establishing their species specific organization. Two types of transcripts have been reported independently in GenBank for human and cattle

Smoc-1 (Figure 57) but their detailed characterization were not reported. In this study, we confirmed presence of two variants of this gene varying in their 3'UTR lengths.

This may either be due to the presence of an alternative splice site within the possible inserted intron (12^{th}) in the 3' region or an alternate splice site in the existing intron (11^{th}) within the 3'UTR itself. However, first possibility seems to be invalid since end point PCR conducted with buffalo genomic DNA using primers from exon 11 and 12 gave rise to a single band of the similar size as that with cDNA. Further, analysis has shown that both the variants have polyadenylation signals 30 and 16 nucleotides upstream to the poly(A) tail for variants -01 & -02 respectively. This is in agreement with the fact that the signals are most often present at 11-30 nucleotides upstream from the poly(A) tail (Sachs, 1993). However, presence of more copies of mRNA instability motifs in variant-01, compared to that in variant -02 supports relatively higher expression of the latter since these motifs are responsible for the degradation of mRNA molecule (Tatrai *et al.*, 2006).

5.2.2.4 *The single copy Smoc-1 with highest expression in liver*

Buffalo *Smoc*-1 is a single copy number gene, and thus, the presence of two variants of this gene may signify either for a backup of the transcripts if one is degraded/mutated or for the enhanced protein expression. In earlier studies, *Smoc*-1 mRNA was reported to be ubiquitously present in all the tissues of mice, showing abundance in ovary but negligible expression in liver and other tissues (Vannahme *et al.*, 2003). Contrary to this, buffalo liver was enriched with *Smoc*-1 transcripts as well as protein whereas other tissues contained fewer or no transcript/ protein substantiating species and tissue specificity of this gene (Figure 69). Liver is primarily involved in vascular functions, metabolic regulation and secretory and excretory functions. Role of the other basement membrane proteins like agrin, collagen IV, laminin and fibronectin in liver cirrhosis and hepatocellular carcinoma has been studied (Wrobel *et al.*, 1979) but no report is available on the functional attributes of *Smoc*-1 in liver. Due to its possible involvement in cell proliferation, adhesion and

tissue remodeling, *Smoc*-1 may also play a pivotal role in hepatocellular activities. Buffalo may not be prone to hepatocellular carcinoma. Since, *Smoc*-1 is conserved across the species, it may be appropriate to study the expression of this gene in human hepatocellular carcinoma to ascertain its possible up- or down regulation.

5.2.2.5 *Smoc-1 and age specific expression: an anticipated view*

Previous studies have shown that *Smoc*-1 mRNA is synthesized even during the early stages of mouse embryonic development. During the embryonic stage day 12, and fetal stage days 14, 16, and 18, the protein is present in the basement membrane zones of various tissues like brain, skin, skeletal muscle, liver, kidney etc (Gersdorff *et al.*, 2006). But, so far no report is available on the sustenance of expression of *Smoc*-1 during life-span of any of the species. Our work seems to be the first report showing a remarkable rise in the *Smoc*-1 expression during 10-14 months of age in buffaloes, followed by constant level maintained throughout their entire life span (Figure 71). As the *Smoc*-1 is supposed to be involved in cell-matrix interaction and bone mineralization, its fulminant expression at 10 month of age and beyond signifies its requirement for growth, development and possible sustenance of the animal.

5.2.2.6 *Smoc-1: tissue localization and future aspects*

Smoc-1 has been localized in zona pellucida and extracellular matrix of mouse ovary. This was suggested to be crucial not only for survival of the oocyte but also for successful fertilization (Gersdorff *et al.*, 2006). In this study, the Smoc-1 has been localized to the extracellular matrix and in the epithelial basement membrane zone of buffalo liver (Figure 72). In addition, staining around the seminiferous tubules and sertoli cells of testis substantiated the true basement membrane localization of Smoc-1 protein (Figure 73) because the basal lamina of the seminiferous tubules in bovines is multilayered and possesses knob like protrusions (Wrobel *et al.*, 1979).

The liver contains a unique extracellular matrix (ECM) within the space of Disse, which consists of basement membrane constituents as

well as fibrillar ECM molecules. Though the basement membranes are mainly formed by a collagen IV, Laminin-1, and nidogen-1 network (Timple, 1996), the liver derived basement membrane also contains a unique isoform composition of type IV collagen, known to bind with the Smoc-1 protein (Zeisberg *et al.*, 2006). Thus, Smoc-1 in ECM of buffalo liver seems to have significance. Changes in the composition of ECM may be detrimental for the viability of hepatocytes during progression of liver cirrhosis. The role of SPARC/Osteonectin in human hepatocellular carcinoma has been reported (Le Bail *et al.*, 1999). Owing to its conservation in human and non-human systems, the fate of *Smoc*-1 gene may be studied in human liver cirrhosis, hepatocellular carcinoma and other liver infections to highlight its clinical aspects.

CONCLUSIONS

6. Conclusions

Present study deals with the identification and characterization of the mRNA transcripts tagged with the simple repeats of the GACA, GATA and 33.15 repeats in water buffalo as a model system, which unveiled the differential organization and expression of these transcripts among different tissues and spermatozoa. Moreover, the detailed isolation and characterization of the full length *Smoc*-1 and *c-kit* genes were also performed to gain insight into their structural and functional organization, expressional status and chromosomal localization. The following points summarize the outcome of this work highlighting the novel potential implications of the MASA uncovered mRNA transcripts in the spermatogenesis, fertilization and several other regulatory pathways:

1. The *in-silico* distributional analyses of the GACA and GATA repeats in the coding and non-coding genomes of Archeas and 17 eukaryotes revealed total absence of these repeats in the prokaryotes, and their accumulation in the higher eukaryotes with an increase of their genetic complexities during evolution. This highlights the significance of these simple repeats in the genome evolution.
2. The analysis of chromosome-wise distribution for the GACA/GATA repeats highlights their preferential accumulation on the mammalian sex chromosomes which suggests their involutions in functional regulation of the sex determination.
3. MASA using the GACA, GATA and 33.15 repeats uncovered a total of 616 fragments, encompassing 148 with 33.15 repeat, 332 with GACA, and 136 with GATA, from somatic tissues, gonads and spermatozoa. This is a novel attempt revealing the existence of the 33.15, GACA and GATA-tagged transcripts in the buffalo spermatozoa highlighting their involvements during the pre- and post-fertilization events.
4. Characterization of the MASA uncovered fragments led to the identification of a total of 63 different mRNA transcripts (34, GACA-tagged; 10, GATA-tagged; and 19, 33.15-tagged) in water buffalo. This association of repeats with the mRNA transcripts, which can be study acquires considerable significance since it establishes the extrapolated

to establish the conclusive significance of other repeats within and adjacent to the coding regions.

5. Exclusive presence of several GACA-/ 33.15-tagged transcripts in a tissue or spermatozoa, and absence of the GATA-tagged transcripts in lung/heart highlights their differential transcript profiles. Several tissue-specific transcripts demonstrate their exclusive requirement in that tissue and their absence showed the transcriptional quiescence in other tissues.
6. The homology search established the novel status of about 50% of the GACA-/33.15-tagged and all the GATA-tagged transcripts corroborating their species-specific distribution. However, the transcripts showing homology to characterized genes were either involved in signal transduction or cell-cell interaction pathways indicating their crucial roles in various cellular functions essential for the life cycle of a cell.
7. Present study also established the GACA richness of the buffalo transcriptome while other species including human were observed to be GATA rich. The GC-richness seems to be unique for the buffalo genome organization and thus for replication timings, methylation and gene expression.
8. All these uncovered mRNA transcripts showed faithful evolutionary conservation across thirteen different species, suggesting broader significance of the 33.15 and GACA/GATA repeats and their tagged transcripts in eukaryotes.
9. Of all the transcripts, approximately 35% demonstrated inter-tissue and/or tissue-spermatozoal sequence polymorphisms which were confirmed in 5-10 additional animals. This may be explained either towards their various functions in different tissues and spermatozoa, or differential functions at various stages of development.
10. The quantitative expressional studies demonstrated the uniform expression of about 30% GACA-tagged in all the sources, whereas 10% GACA-tagged and 15% 33.15 tagged transcripts with highest expression in the liver or spleen indicated their putative involvement in the hepatocellular and immunological activities, respectively.

11. Most interestingly, the exclusive or highest expression of 60% GACA-tagged, 85% 33.15-tagged, and 100% GATA-tagged transcripts in the testis and/or spermatozoa substantiating their deep involutions in various testicular functions like spermatogenesis and male gonad development.

12. The full length CDS of proto-oncogene *c-kit* (2973 bp) from different tissues, and Secreted modular calcium binding protein-1 (3474 bp) from liver, of water buffalo were isolated.

13. Upon comparison, the *c-kit* sequences showed tissue-specific nucleotide insertions, deletions and changes resulting in novel truncated peptides. These peptides lacked intracellular and/or transmembrane domains in all other tissues. However, only testis was found to encode full length c-kit protein which highlighted its tissue and stage specific functions.

14. *C-kit*, implicated with spermatogenesis, melanogenesis and hematopoeisis, was found to undergo tissue-specific alternate splicing. These alternately spliced transcripts were the integral parts of the open reading frame and have been reported in other mammals.

15. Multiple sequence alignment of *c-kit* sequences across the mammals revealed a unique tyrosine kinase domain in buffalo compared to that in other species, suggesting the species-specific organization and function of *c-kit*.

16. The expressional analysis of *c-kit* unveiled its highest expression in testis, and 10 times lesser in spermatozoa compared to that in testis which substantiates its predominant role in spermatogenesis. This study establishes unequivocal involvement of an autosomal gene *c-kit* receptor in testicular functions.

17. Present study demonstrates the association of the consensus sequence of minisatellite 33.15 with the *Smoc*-1 which was found to encode a secreted matricellular glycoprotein containing two EF-hand calcium binding motifs homologous to that of BM-40/SPARC family which suggest their property of calcium binding.

18. This gene consisting of 12 exons was mapped onto the acrocentric chromosome 11 in buffalo. Though this gene was found to be

evolutionarily conserved, the buffalo *Smoc*-1 showed conspicuous nucleotide/amino acid changes altering its secondary structure compared to that in other mammals. This suggests their species-specific organizational and functional uniqueness.

19. Two EF-hand motifs in the ECD conformed well to its calcium binding affinity and N-glycosylation site at Asn-214 suggesting its glycoprotein nature with a calcium dependent conformation.

20. For the first time, we unveiled two transcript variants of this gene, varying in their 3'UTR lengths but both coding for identical protein(s). Buffalo *Smoc*-1 is a single copy number gene, and thus, the presence of two variants of this gene may signify either for a backup of the transcripts if one is degraded/mutated or for the enhanced protein expression.

21. Buffalo *Smoc*-1 evidenced highest expression of both the variants in liver and modest to negligible in other tissues contrary to the earlier reports in mouse substantiating species and tissue-specific functions of this gene. The relative expression of variant-02 was markedly higher compared to that of variant-01 in all the tissues examined, highlighting the variant-02 as major transcript and variant-01 as minor one.

22. Moreover, the expression of Smoc-1, though moderate during the early ages, was conspicuously enhanced after 1 year age and remained consistently higher during the entire life span of buffalo, with the gradual increment in expression of variant-02, intimating its role in postnatal development besides embryonic development. Since *Smoc*-1 is thought to be involved in cell-matrix interaction and bone mineralization, its fulminant expression at 10 month age and beyond which signifies its requirement for growth, development and possible sustenance of the animal.

REFERENCES

7. REFERENCES

1. Aharoni, A., Baran, N. and Manor, H. (1993). Characterization of a multisubunit human protein which selectively binds single stranded d(GA) n and d (GT) n sequence repeats in DNA. *Nucl. Acids Res*, **21**: 5221-5228.
2. Ali, S. and Wallace, R.B. (1988). Intrinsic polymorphism of variable number tandem repeat loci in the human genome. *Nucl. Acids Res.* **16**: 8487-8496.
3. Ali, S., Azfer, A.A., Bashamboo, A., Mathur, P.K., Malik, P.K., Mathur, V.B., Raha, A.K., and Ansari, S. (1999). Characterization of a species-specific repetitive DNA from a highly endangered wild animal, *Rhinoceros unicornis*, and assessment of genetic polymorphism by microsatellite associated sequence amplification (MASA). *Gene* **228**: 33-42.
4. Alliel, P.M., Perin, J.P., Jolles, P. and Bonnet, F.J. (1993) Testican, a multidomain testicular proteoglycan resembling modulators of cell social behaviour. *Eur. J. Biochem.* **214:** 347–350.
5. Andre, C., Martin, E., Cornu, F., Hu, W., Wang, X. and Gilbert, F. (1992). Genomic organization of the human *c-kit* gene: evolution of the receptor tyrosine kinase subclass III. *Oncogene* **7**: 685-691.
6. Areshchenkova, T. and Ganal, M.W. (1999) Long tomato microsatellites are predominantly associated with centromeric regions. *Genome* **42**: 536-544.
7. Armour, J.A.L. (1999). Microsatellites and mutation processes in tandemly repetitive DNA. In: Goldstein, D. and Schlotterer, C. Editors, *Microsatellites: Evolution and Applications*, *Oxford University Press, Oxford* 24–29.
8. Baldi, P. and Basnee, P.F. (2000) Sequence analysis by additive scale: DNA structure for sequences and repeats of all lengths. *Bioinformatics*, **16**: 865-889.
9. Bashamboo, A. and Ali, S. (2001). Minisatellite associated sequence amplification (MASA) of the hypervariable repeat marker

33.15 reveals a male specific band in humans. *Mol. Cell. Probes* **15**: 89-92.

10. Bennett, P. (2000). Microsatellites. *J. Clin. Pathol.* **53**: 177-183.
11. Bertoni, F., Codegoni, A.M., Furlan, D., Tibiletti, M.G., Capella, C. and Broggini, M. (1999). CHK1 frameshift mutations in genetically unstable colorectal and endometrial cancers. *Genes Chromosomes Canc.* **26**: 176-180.
12. Besmer, P., Murphy, P.C., George, P.C., Bergold, P.T., Lederman, L., Synder, H.W., Brodeur, D., Zukkerman, E.E. and Hardy, W.D. (1986). A new acute transforming feline retrovirus and relationship of its oncogene v-kit with the protein kinase family. Nature **320**, 415-421.
13. Besmer, P. (1995). Differential roles of I-3 kinase and kit 821 in kit receptor mediated proliferation, survival and cell adhesion in mast cells. *EMBO J.* **12**: 473-483.
14. Biet, E., Sun, J. and Dutreix, M. (1999). Conserved sequence preference in DNA binding among recombination proteins: an effect of ssDNA secondary structure. *Nucl. Acids Res.* **27**: 596-600.
15. Bonaccorsi, S., Gatti, M., Pisano, C. and Lohe, A. (1990). Transcription of a satellite DNA on two Y chromosome loops of *Drosophila*. *Chromosoma* **99**: 260–266.
16. Borstnik, B., Pumpernik, D., Lukman, D., Ugarkovic, D. and Plohl, M. (1994). Tandemly repeated pentanucleotides in DNA sequences of eukaryotes. *Nucl. Acids Res.* **22**: 3412–3417
17. Borstnik, B. and Pumpernik, D. (2002). Tandem repeats in protein coding regions of primate genes. *Genome Res.* **12**: 909-915.
18. Brandes, A., Thompson, H., Dean, C. and Heslop-Harrison, J.S. (1997). Multiple repetitive DNA sequences in the paracentromeric regions of *Arabidosis thaliana* L. *Chromosome Res.* **5**: 238-246.
19. Brekken, R.A. and Sage, E.H. (2001). SPARC, a matricellular protein, at the crossroads of cell-matrix communication. *Matrix Biol.* **19**: 816-827.
20. Brock, G.J., Anderson, N.H. and Monckton, D.G. (1999). Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability:

association with flanking GC content and proximity to CpG islands. *Hum. Mol. Genet.* **8**: 1061-1067.

21. Capy, P., Bazin, C., Higuet, D. and Langin, T. (1998). Dynamics and evolution of transposable elements. *Landes Bioscience Austin, Texas.*
22. Carim-Todd, L., Escarceller, M., Estivill, X. and Sumoy, L. (2003). LRRN6A/LERN1 (leucine-rich repeat neuronal protein 1), a novel gene with enriched expression in limbic system and neocortex. *Eur. J. Neurosci.* **18(12)**: 3167-3182.
23. Carrol, S.B. (2000) Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**: 577–580.
24. Catasti, P., Chen, X., Mariappan, S.V., Bradbury, E.M. and Gupta, G. (1999) DNA repeats in the human genome. *Genetica* **106**: 15-36.
25. Chabot, B., Stephenson, D.A., Chapman, V.M., Besmer, P. and Bernstein, A. (1988). The protooncogene *c-kit* encoding a tyrosine kinase receptor maps to the mouse W locus. *Nature* **335**: 88-89.
26. Chamberlain, N.L., Driver, E.D. and Miesfeld, R.L. (1994). The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. *Nucl. Acids Res.* **22**: 3181-3186.
27. Chambers, G.K. and MacAvoy, E.S. (2000). Microsatellites: consensus and controversy. *Comp. Biochem. Physiol. B* **126**: 455–476.
28. Chang, D.K., Metzgar, D., Wills, C. and Boland, C.R. (2001). Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. *Genome Res.* **11**: 1145-1146.
29. Charlesworth, B., Sniegowski, P. and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**: 215-220.
30. Chiodi, I., Corioni, M., Giordano, M., Valgardsdottir, R., Ghigna, C., Cobianchi, F., Xu, R.M., Riva, S. and Biamonti, G. (2004). RNA recognition motif 2 directs the recruitment of SF2/ASF to nuclear stress bodies. *Nucl. Acids Res.* **32**: 4127-4136.

31. Codegoni, A.M., Bertoni, F., Collele, G., Grassi, L., D'Incalci, M. and Broggini, M. (1999). Microsatellite instability and frameshift mutations in genes involved in cell cycle progression or apoptosis in ovarian cancer. *Oncol. Res.* **11**: 297-301.
32. Cohen, H., Danin-Poleg, Y., Cohen, C.J., Sprecher, E., Darvasi, A. and Kashi, Y. (2004). Mono-nucleotide repeats (MNRs): a neglected polymorphism for generating high density genetic maps in silico. *Hum. Genet.* **115**: 213–220.
33. Cooper, J.L. and Henikoff, S. (2004). Adaptive evolution of the histone fold domain in centromeric histones. *Mol. Biol. Evol.* **21**: 1712–1718.
34. Crosier, P.S., Ricciardi, S.T., Hall, L.R., Vitas, M.R., Clark, S.C. and Crosier, S.E. (1993). Expression of isoforms of the human receptor tyrosine kinase *c-kit* in leukemic cell lines and acute myeloid leukemia. *Blood* **82**: 1151-1158.
35. Csink, A.K. and Henikoff, S. (1998). Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.* **14**: 200–204.
36. Cuadrado, A. and Schwarzacher, T. (1998). The chromosomal organization of simple sequence repeats in wheat and rye genomes. *Chromosoma* **107**: 587-594.
37. Cummings, C.J. and Zoghbi, H.Y. (2000). Trinucleotide repeats: mechanisms and path physiology. *Annu. Rev. Genomics Hum. Genet.* **1**:281–328.
38. Davis, B.M., McCurrach, M.E., Taneja, K.L., Singer, R.H. and Housman, D.E. (1997). Expansion of a CUG trinucleotide repeat in the 39 untranslated regions of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts. *Proc. Natl. Acad. Sci. USA*. **94**: 7388–7393.
39. Di Prospero, N.A. and Fischbeck, K.A (2005). Therapeutic development for triplet repeat expansion diseases. *Nat. Rev. Genet.* **6**: 756-765.
40. Diaz, M.O., Barsacchi-Pilone, G., Mahon, K.A., Gall, J.G. (1981). Transcripts from both DNA strands of a satellite DNA occur on

lampbrush chromosome loops of the newt *Notophthalmus*. *Cell* **24**: 649–659.

41. Dieringer, D. and Schlotterer, C. (2003). Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res.* **13**:2242–2251.
42. Dokholyan, N.V., Buldyrev, S.V., Havlin, S. and Stanley, H.E. (2000). Distributions of dimeric tandem repeats in noncoding and coding DNA sequences. *J. Theor. Biol.* **202**: 273-282.
43. Dover, G.A. (1986). Molecular drive in multigene families: How biological novelties arise, spread and are assimilated. *Trends Genet.* **2:** 159–165.
44. Duret, L., Semon, M., Piganeaue, G., Mouchiroud, D. and Galtier, N. (2002). Vanishing GC-rich Isochores in Mammalian genomes. *Genetics* **162**: 837-1847.
45. Dushlaine, C.T.O., Edwards, R.J., Park, S.D. and Shields, D.C. (2005). Tandem repeat copy number variation in protein-coding regions of the human genes. *Genome Biol.* **6**: R69.
46. Dutreix, M. (1997). (GT)n repetitive tracts affect several stages of RecA-promoted recombination. *J. Mol. Biol.* **273**: 105-113.
47. Edwards, Y.J., Elgar, G., Clark, M.S. and Bishop, M.J. (1998). The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, Fugu rubripes: perspectives in functional and comparative genomic analyses. *J. Mol. Biol.* **278**: 843–854.
48. Eib, D.W and Martens, G.J. (1996). A novel transmembrane protein with epidermal growth factor and follistatin domains expressed in the hypothalamo-hypophysical axis of Xenopus laevis. *J. Neurochem.* **67**: 1047-1055.
49. Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* **5**: 435–445.
50. Epplen, J.T., McCarrey, J.R., Sutou, S. and Ohno, S. (1982). Base sequence of a cloned snake W-chromosome DNA fragment and identification of a male-specific putative mRNA in the mouse. *Proc. Natl. Acad. Sci. USA* **79**: 3798-3802.

51. Epplen, J.T. (1988). On simple repeated GATCA sequences in animal genomes: a critical reappraisal. *J. Hered.* **79(6)**: 409-417.

52. Fabre, E., Dujon, B and Richard, G.F. (2002). Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. *Nucl. Acids Res.* **30**:3540–3547.

53. Fabregat, I., Koch, K.S., Aoki, T., Atkinson, A.E., Dang, H., Amosova, O., Fresco, J.R., Schidkraut, C.L. and Leffert, H.L. (2001). Functional pleiotropy of an intramolecular triplex-forming fragment from the 3-UTR of the rat *Pigr* gene. *Physiol. Genomics* **5**: 53-65.

54. Field D. and Christopher Wills, C. (1998). Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, results from strong mutation pressures and a variety of selective forces. *Proc. Natl. Acad. Sci. USA* **95**:1647–1652.

55. Fitzgerald, M. and Shenk, T. (1981). The sequence 5'-AAUAAA-3'forms parts of the recognition site for polyadenylation of late SV40 mRNAs. *Cell* **24**: 251-260.

56. Gacy, A.M., Goellner, G., Juranic, N., Macura, S. and McMurray, C.T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. *Cell* **81**: 533–540.

57. Galli, S.J., Zsebo, K.M. and Geissler, E.N. (1994). The kit ligand, stem cell factor. *Adv. Immunol.* **55**: 1-96.

58. Gangadharan, S., Kapur, V. and Ali, S. (2001). GATA/GACA repeat sequences are transcribed in the normal fertile rat *Rattus norvegicus*, but not in the infertile ones. *Curr. Sci.* **81(10)**: 1320-1324.

59. Gebhardt, F., Zanker, K.S. and Brandt, B. (1999). Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J. Biol. Chem.* **274**: 13176-13180.

60. Gebhardt, F., Burger, H. and Brandt, B. (2000). Modulation of EGFR gene transcription by a polymorphic repetitive sequence Ñ a

link between genetics and epigenetics. *Int. J. Biol. Mark.* **15**: 105-110.

61. Gersdorff, N., Muller, M., Schall, A. and Miosge, N. (2006). Secreted modular calcium-binding protein1 localization during mouse embryogenesis. *Histochem. Cell. Biol.* **126(6)**: 705-712.
62. Girard, J.P. and Springer, T.A. (1995). Cloning from purified high endothelial venule cells of hevin, a close relative of the antiadhesive extracellular matrix protein SPARC. *Immunity* **2:** 113-123.
63. Gokkel, E., Grossman, Z., Ramot, B., Yarden, Y., Rechavi, G. and Givol, D. (1992). Structural organization of the murine *c-kit* Protooncogene. *Oncogene* **7**: 1423-1429.
64. Guermah, M., Crisanti, P., Laugier, D., Dezelee, P., Bidou, L., Pessac, B. and Calothy, G. (1991). Transcription of a quail gene expressed in embryonic retinal cells is shut off sharply at hatching. *Proc. Natl. Acad. Sci. USA* **88:** 4503–4507.
65. Hall, S.E., Kettler, G. and Preuss, D. (2003). Centromere satellites from *Arabidopsis* populations: maintenance of conserved and variable domains. *Genome Res.* **13**: 195–205.
66. Hancock, J.M. (1996). Simple sequences and the expanding genome. *Bioessays* **18**: 421-425.
67. Harding, R.M., Boyce, A.J. and Clegg, J.B. (1992). The evolution of tandemly repetitive DNA: recombination rules. *Genetics* **132**: 847-859.
68. Hefferon, T.W. (2004). Avariable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. *Proc. Natl. Acad. Sci. USA* **101**: 3504–3509.
69. Henaut, A., Lisacek, F., Nitschke, P., Moszer, I. and Danchin, A. (1998). Global analysis of genomic texts: the distribution of AGCT tetranucleotides in the *Escherichia coli and Bacillus subtilis* genomes predicts translational frameshifting and ribosomal hopping in several genes. *Electrophoresis* **19:** 515-527.
70. Henikoff, S. and Dalal, Y. (2005). Centromeric heterochromatin: what makes it unique? *Curr. Opin. Genet. Dev.* **15**: 177–184.

71. Hobza, R., Lengerova, M., Svoboda, J., Kubekova, H., Kejnovsky. E. and Vyskot, B. (2006). An accumulation of tandem DNA repeats on the Y chromosome in *Silene latifolia* during early stages of sex chromosome evolution. *Chromosoma* **115(5)**: 376-382.

72. Hoffman, E.K., Trusko, S.P., Murphy, M. and George, D.L. (1990). An S1 nuclease-sensitive homopurine/homopyrimidine domain in the c-Ki-ras promoter interacts with a nuclear factor. *Proc. Natl. Acad. Sci. USA* **87**: 2705-2709.

73. Hohenester, E., Maurer, P., Hohenadl, C., Timpl, R., Jansonius, J.N. and Engel, J. (1996). Structure of a novel extracellular Ca^{2+}-binding module in BM-40. *Nat. Struct. Biol.* **3**: 67–73.

74. Hood, D.W., Deadman, M.E., Jennings, M.P., Bisercic, M., Fleischmann, R.D., Venter, J.C. and Moxon, E.R. (1996). DNA repeats identify novel virulence genes in Haemophilus influenzae. *Proc. Natl. Acad. Sci. USA* **93**:11121–11125.

75. Horie, M., Mitsumoto, Y., Kyushiki, H., Kanemoto, N., Watanabe, A., Taniguchi, Y., Nischino, N., Okamoto, T., Kondo, M., Mori, T. *et al.* (2000). Identification of characterization of TMEFF2, a novel survival factor for hippocampel and mesencephalic and neurons. *Genomics* **67**: 146-152.

76. Ivanov, I., Alexandrova, R., Dragulev, B., Saraffova, A. and Abou Haidar, A.G. (1992). Effect of tandemly repeated AGG triplets on the translation of CAT-mRNA in *E. coli*. *FEBS Lett.* **307**: 173-176.

77. Jasinska, A. and Krzyzosiak, W.J. (2004). Repetitive sequences that shape the human transcriptome. *FEBS lett.* **567**: 136-141.

78. Jeffreys, A.J., Wilson, V. and Stein, S.L. (1985). Hypervariable minisatellite" regions in human DNA. *Nature* **314**: 67–73.

79. Jeffreys, A.J., Royle, N.J., Wilson, V. and Wong, Z. (1998). Spontaneous mutation rates to new length alleles at tandem-repetitive hyper-variable loci in human DNA. *Nature* **332(6161)**: 278-281.

80. Jeffreys, A.J., Murray, J. and Neumann, R. (1998). High-resolution mapping of crossovers in human sperm defines a minisatellite associated recombination hotspot. *Mol. Cell* **2:** 267-273.

81. Johannsdottir, J.T., Jonasson, J.G., Bergthorsson, J.T., Amundadottir, L.T., Magnusson, J., Egilsson, V. and Ingvarsson, S. (2000). The effect of mismatch repair deficiency on tumourigenesis; microsatellite instability affecting genes containing short repeated sequences. *Int. J. Oncol.* **16**: 133-139.

82. John, M.V., and Ali, S. (1997). Synthetic DNA-based genetic markers reveal intra- and inter-species DNA sequence variability in the *Bubalus bubalis* and related genomes. *DNA Cell Biol.* **16(3)**: 369-378.

83. Jordan, P., Snyder, L.A.S. and Saunders, N.J. (2003). Diversity in coding tandem repeats in related *Neisseria* spp. *BMC Microbiol.* **3**: 23-37.

84. Kapur, V., Prasanth, S.G., O'Ryan, C., Md. Azfer, A. and Ali, S. (2003). Development of a DNA marker by minisatellite associated sequence amplification (MASA) from the endangered Indian Rhino (Rhinoceros unicornis. *Mol. Cell. Probes* **17**: 1-4.

85. Karlin, S., Mrazek, J. and Campbell, A.M. (1997). Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179**:3899–3913.

86. Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J. and Gentles, A.J. (2002). Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. USA* **99**: 333–338.

87. Kashi, Y., King, D.G. and Soller, M. (1997). Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* **13**: 74-78.

88. Kashi, Y. and King, D.G. (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* **22**: 253-259.

89. Katti, M.V., Prabhakar, K., Ranjekar and Gupta, V.S. (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18**: 1161-1167.

90. Kierszenbaum, A.L. (2006). Tyrosine protein kinases and spermatogenesis: truncation matters. *Mol. Reprod. Dev.* **73**: 399-403.

91. King, D.G. and Soller, M. (1999). Variation and fidelity: The evolution of simple sequence repeats as functional elements in adjustable genes. *Trends Genet.* 22: 253-259.

92. Kizawa, H., Kou, I., Lida, A., Sudo, A., Miyamoto, Y., Fukuda, A., Mabuchi, A., Kotani, A., Kawakami, A., Yamamoto, S. *et al.* (2005). An aspartic acid repeat polymorphism in asporin inhibits chondrogenesis and increases susceptibility to osteoarthritis. *Nat. Genet.* **37(2)**: 138-144.

93. Kliebenstein, D.J., West, M.A.L., Leeuwen, Hans van., Kim, K., Doerge, R.W., Michelmore, R.W. and Clair, D.A. (2006). Genomic Survey of Gene Expression Diversity in *Arabidopsis thaliana. Genetics* **172(2)**: 1179-1189.

94. Kolodner, R.D. and Marsischky, G.T. (1999). Eukaryotic DNA mismatch repair. *Curr. Opin. Genet. Dev.* **9**: 89-96.

95. Krawetz, S.A. (2005). Paternal contribution: new insights and future challenges. *Nat. Rev. Genet.* **6**: 633-642.

96. Lambard, S., Galeraud-Denis, I., Martin, G., Levy, R., Chocat, A. and Carreau, S. (2004). Analysis and significance of mRNA in human ejaculated sperm from normozoospermic donors: relationship to sperm motility and capacitation. *Mol. Hum. Reprod.* **10(7)**: 535-541.

97. Lane, T.F. and Sage, E.H. (994) The biology of SPARC, a protein that modulates cell-matrix interactions. *FASEB J.* **8**: 163-173.

98. Le Bail, B., Faouzi, S., Boussarie, L., Guirouilh, J., Blanc, J.F., Carles, J., Bioulac-Sage, P., Balabaud, C. and Rosenbaum, J. (1999). Osteonectin/SPARC is over-expressed in human hepatocellular carcinoma. *J. Pathol.* **189(1)**: 46-52.

99. Levinson, G. and Gutman, G.A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution, *Mol. Biol. Evol.* **4**: 203–221.

100. Li, Y.X. and Kirby, M.L. (2003). Coordinated and conserved expression of alphoid repeat and alphoid repeat-tagged coding sequences. *Dev. Dyn.* **228**: 72–81.

101. Li, Y.C., Korol, A.B., Fahima, T., Nevo, E. (2004). Microsatellites within genes: structure, function and evolution. *Mol. Biol. Evol.* **21**: 991-1007.

102. Liu, Z., Tan, G., Li, P. and Dunham, R.A. (1999). Transcribed dinucleotide microsatellites and their associated genes from channel catfish *Ictalurus punctatus*. *Biochem. Biophys. Res. Commun.* **259**: 190-194.

103. Liu, L., Dybvig, K., Panangala, V.S., van Santen, V.L., French, C.T. (2000). GAA trinucleotide repeat region regulates M9/pMGA gene expression in *Mycoplasma gallisepticum*. *Infec. Immun.* **68**: 871-876.

104. Luconi, M., Marra, F., Gandini, L., Filimberti, E., Lenzi, A., Forti, G. and Baldi, E. (2001). Phosphatidylinositol 3-kinase inhibition enhances human sperm motility. *Hum. Reprod.* **16(9)**: 1931-1937.

105. Majewski, J. and Ott, J. (2000). GT repeats are associated with recombination on human chromosome 22. *Genome Res.* **10**: 1108-1114.

106. Marcotte, E.M., Pellegrini, M., Yeates, T.O. and Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* **293**: 151-160.

107. Maurer, P., Hohenadl, C., Hohenester, E., Gohring, W., Timpl, R. and Engel, J. (1995). The C-terminal portion of BM-40 (SPARC/osteonectin) is an autonomously folding and crystallisable domain that binds calcium and collagen IV. *J. Mol. Biol.* **253**: 347–357.

108. McDonald, J.F. (1995). Transposable elements: possible catalysis of organismic evolution. *Trends Ecol. Evol.* **10**:123-126.

109. Meloni, R., Albanese, V., Ravassard, P., Treilhou, F. and Mallet, J. (1998). A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element *in vitro*. *Hum. Mol. Genet.* **7**: 423-428.

110. Metz, A., Soret, J., Vourch, C., Tazi, J. and Jolly, C. (2004). A key role for stress-induced satellite III transcripts in the relocalization of splicing factors into nuclear stress granules. *J. Cell. Sci.* **117**: 4551–4558.

111. Metzgar, D., Bytof, J. and Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.* **10**:72–80.

112. Miller, D., Tang, P.Z., Skinner, C. and Lilford, R. (1994). Differential RNA fingerprinting as a tool in the analysis of spermatozoal gene expression. *Hum. Reprod.* **9(5)**: 864-869.

113. Miller, D. (2000). Analysis and significance of messenger RNA in human ejaculated spermatozoa. *Mol. Reprod. Dev.* **56**: 259-264.

114. Miller, D., Ostermeier, G.C. and Krawetz, S.A. (2005). The controversy, potential and roles of spermatozoal RNA. *Trends Mol. Med.* **11(4)**: 156-163.

115. Molina, F., Bouanani, M., Pau, B. and Granier, C. (1996). Characterization of the type-1 repeat from thyroglobulin, a cysteine-rich module found in proteins from different families. *Eur. J. Biochem.* **240**: 125-133.

116. Morales, P.J., Vantman, D., Barros, C. and Vigil, P. (1991). Human spermatozoa selected by Percoll gradient or swim-up are equally capable of binding to the human zona pellucida and undergoing the acrosome reaction. *Hum. Reprod.* **6**: 401-404.

117. Morgante, M., Hanafey, M., and Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30**: 194–200.

118. Mosavi, L.K., Cammet, T.J., Desrosiers, D.C. and Peng, Z.Y. (2004). The ankyrin repeat as molecular architecture for protein recognition. *Prot. Sci.* **13(6)**: 1435-1438.

119. Moxon, E.R., Rainey, P.B., Nowak, M.A., Lenski, R.E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**: 24-33.

120. Mravinac, B., Plohl, M., Ugarkovic, D. (2005). Preservation and high sequence conservation of satellite DNAs indicate functional constraints. *J. Mol. Evol.* **61**: 542-550.

121. Murphy, T.D. and Karpen, G.H. (1995). Localization of centromere function in a Drosophila minichromosome. *Cell* **82(4):** 599-609.

122. Murphy, G.L., Connell, T.D., Barritt, D.S., Koomey, M. and Cannon, J.G. (1989). Phase variation of gonococcal protein II: regulation of gene expression by slipped strand mispairing of a repetitive DNA sequences. *Cell* **56**:539–547.

123. Nadir, E., Hargalit, H., Gallily, T. and Ben-Sasson, S.A. (1996). Microsatellite spreading in the human genome: Evolutionary mechanisms and structural implications. *Proc. Natl. Acad. Sci. USA* **93**: 6470-6475.

124. Nakada, M., Yamada, A., Takimo, T., Miyamori, H., Takahashi, T., Yamashita, J. and Sato, H. (2001). Suppression of membrane-type 1 matrix metalloproteinase (MMP)-mediated MMP-2 activation and tumor invasion by testican 3 and its splicing variant gene product, N-Tes. *Canc. Res.* **61**: 8896-8902.

125. Nakamura, Y., Leppert, M., O'Connel, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. *et al.* (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**: 1616-1622.

126. Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Pierre-Edouard, F., Raoult, D. and Jean-Michel, C. (2000). Selfish DNA in Protein-Coding Genes of Rickettsia. *Science* **290**: 347.

127. Okladnova, O., Syagailo, Y.V., Tranitz, M., Stober, G., Riederer, P., Mossner, R., and Klaus-Peter, L. (1998). A promoter associated polymorphic repeat modulates PAX-6 expression in human brain. *Biochem. . Biophys. Res. Comm.* **248**: 402-405.

128. Ostermeier, G.C., Goodrich, R.J., Moldenhauer, J.S., Diamond, M.P. and Krawetz, S.A. (2005). A suite of novel human spermatozoal RNAs. *J. Androl.* **26(1)**: 70-74.

129. Paronetto, M.P., Farini, D., Sammarco, I., Maturo, G., Vaspasiani, G., Geremia, R., Rossi, P. and Sette, C. (2004). Expression of a truncated form of the *c-kit* tyrosine kinase receptor and activation of Src kinase in human prostate cancer. *Am. J. Pathol.* **164**: 1243-1251.

130. Petitpierre, E., Juan, C., Pons, J., Plohl, M, and Ugarkovi, D. (1995). Satellite DNA and constitutive heterochromatin in

tenebrionid beetles. In Brandham, P.E. and Bennett,M.D. (eds), Kew Chromosome Conference IV. *Royal Botanic Gardens, London*, 351–362.

131. Prasanth, S.G., Giran, H.M. and Ali, S. (2004). Biology of proto-oncogene *c-kit* receptor in spermatogenesis. *Curr. Pharmacogen.* **2**: 47-60.
132. Qian, J., Yang, J., Zhang, X., Zhang, B., Wang, J., Zhou, M., Tang, K., Li, W., Zeng, Z., Zhao, X., Shen, S. and Li, G. (2001). Isolation and characterization of a novel cDNA, UBAP1, derived from the tumor suppressor locus in human chromosome 9p21-22. *J. Canc. Res. Clin. Oncol.* **127(10)**: 613-618.
133. Qiu, F.H., Ray, P., Brown, K., Barker, P.E., Jhanwar, S., Ruddle, F.H. and Besmer, P. (1988). Primary structure of *c-kit*: relationship with the CSF-1/PDGF receptor for kinase family-oncogenic activation of v-kit involves deletion of extracellular domain and c-terminus. *EMBO J.* **7**: 1003-1011.
134. Reith, A.D., Ellis, C., Lysman, S.D., Anderson, D.M., Williams, D.E., Bernstein, A. and Pawson, T. (1991). Signal transduction by normal isoforms and w mutant variants of the kit receptor tyrosine kinase. *EMBO J.* **10**, 2451-2459.
135. Richards, R.I. (2001). Dynamic mutations: a decade of unstable expanded repeats in human genetic disease. *Hum. Mol. Genet.* **10(20)**: 2187-2194.
136. Rocnik, E.F., Liu, P., Sato, K., Walsh, K. and Vaziri, C. (2006). The novel SPARC family member SMOC-2 potentiates angiogenic growth factor activity. *J. Biol. Chem.* **281(32):** 22855-22864.
137. Robles, F., De La Herran, R., Ludwig, A., Ruiz-REJON, C., Ruiz-Rejon, M., and Garrido-Ramos, M.A. (2004). Evolution of ancient satellite DNAs in sturgeon genomes. *Gene* **338**: 133-142.
138. Rocha, E.P.C., Matric, I. and Taddei, F. (2002). Over-expression of repeats in stress response genes: a strategy to increase versatility under stressful conditions? *Nucl. Acids Res* **30**: 1886-1894.
139. Rockman, M.V. and Wray, G.A. (2002). Abundant raw material for cisregulatory evolution in humans. *Mol. Biol. Evol.* **19**: 1991–2004.

140. Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., Mclay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P. *et al,* (2005). The DNA sequence of the human X chromosome. *Nature* **434(7031)**: 325-337.

141. Rudert, F., Bronner, S., Garnier, J-M. and Dolle, P. (1995). Transcripts from opposite strands of γ satellite DNA are differentially expressed during mouse development. *Mamm. Genome* **6**: 76–83.

142. Sachs, A.B. (1993). Messenger RNA degradation in eukaryotes. *Cell* **174**: 413–421.

143. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). Molecular Cloning I, II and III. *Cold Spring Harbor Laboratory Press.*

144. Sandaltzopoulos, R., Mitchelmore, C., Bonte, E., Wall, G., Becker, P.B. (1995). Dual regulation of the *Drosophila hsp26* promoter *in vitro. Nucl. Acids Res.* **23**: 2479-2487.

145. Sandberg, G., Schalling, M. (1997). Effect of in vitro promoter methylation and CGG repeat expansion on FMR-1 expression. *Nucl Acids Res* **14**: 2883-2887.

146. Saunders, C.M., Larman, G.M., Parrington, J., Cox, L.J., Royse, J., Blayney, L.M., Swann, K. and Lai, F.A. (2002). PLC zeta: a sperm-specific trigger of Ca(2+) oscillations in eggs and embryo development. *Development* **129 (15)**: 3533-3544.

147. Saveliev, A., Everett, C., Sharpe, T., Webster, Z. and Festenstein, R. (2003). DNA triplet repeats mediate heterochromatin-protein-1-sensitive variegated gene silencing. *Nature* **422**:909–913.

148. Schlötterer, C. (2004). The evolution of molecular markers- just a matter of fashion, *Nat. Rev. Genet.* **5**: 63–69.

149. Schmidt, T. and Heslop-Harrison, J.S. (1996). The physical and genomic organization of microsatellites in sugar beet. *Proc. Natl. Acad. Sci. USA* **93**: 8761-8765.

150. Schueler, M.G., Higgins, A.W., Rudd, M.K., Gustashaw, K. and Willard, H.F. (2001). Genomic and genetic definition of a functional human centromere. *Science* **294**: 109–115.

151. Schug, M.D., Wetterstrand, K.A., Gaudette, M.S., Lim, R.H., Hutter, C.H. and Aquadro, C.F. (1998). The distribution and frequency of microsatellite loci in *Drosophila melanogaster*. *Mol. Ecol.* **7**: 57-70.

152. Scotti, I., Vendramin, G.G., Matteotti, L.S., Scarponi, C., Sari-Gorla, M. and Binelli, G. (2000). Postglacial recolonization routes for Picea abies K. in Italy as suggested by the analysis of sequence-characterized amplified region (SCAR) markers. *Mol. Ecol.* **9**: 699-708.

153. Serve, H., Yee, N.S., Stella, G., Sepp-Lorenzino, L., Tan, J.C. and Besmer, P. (1995). Differential roles of I-3 kinase and kit 821 in kit receptor mediated proliferation, survival and cell adhesion in mast cells. *EMBO J.* **12**: 473-483.

154. Shibanuma, M., Mashimo, J., Mita, A., Kuroki, T. and Nose, K. (1993). Cloning from a mouse osteoblastic cell line of a set of transforming-growth-factor-beta 1-regulated genes, one of which seems to encode a follistatin-related polypeptide. *Eur. J. Biochem.* **217:** 13–19.

155. Singh, L., Purdom, I.F. and Jones, K.W. (1980). Sex chromosome associated satellite DNA: evolution and conservation. *Chromosoma* **79(2)**: 137-157.

156. Singh, L. and Jones, K.W. (1982). Sex reversal in the mouse (*Mus musculus*) is caused by a recurrent nonreciprocal crossover involving the x and an aberrant Y chromosome. *Cell* **28(2)**: 205-216.

157. Singh, L., Wadhwa, R., Naidu, S., Nagraj, R. and Gandean, M. (1994). Sex- and tissue-specific Bkm(GATA)-binding protein in the Germ cells of heterogametic sex. *J. Biol. Chem.* **269(41)**: 25321-25327.

158. Srivastava, J., Premi, S., Pathak, D., Ahsan, Z., Tiwari, M., Garg, L.C. and Ali, S. (2006a). Transcriptional status of known and novel genes tagged with consensus of 33.15 repeat loci employing minisatellite-associated sequence amplification (MASA) and real-time PCR in water buffalo, *Bubalus bubalis*. *DNA Cell Biol.* **25(1):** 31-48.

159. Srivastava, J., Premi, S., Garg, L.C. and Ali, S. (2006b). Organizational and expressional uniqueness of a testis-specific mRNA transcript of protooncogene *c-kit* receptor in water buffalo *Bubalus bubalis*. *DNA Cell Biol.* **25(9)**:501-513.

160. Subramanian, S., Mishra, R.K. and Singh, L. (2003). Genome-wide analysis of Bkm sequences (GATA repeats): predominant association with sex chromsosome and potential role in higher order organization and function. *Bioinformatics* **19(6)**: 681-685.

161. Sun, X., Wahlstrom, J. and Karpen, G. (1997). Molecular structure of a functional *Drosophila* centromere. *Cell* **91**: 1007–1019.

162. Sutherland, G.R. and Richards, R.I. (1995). Simple tandem repeats and human genetic disease. *Proc. Natl. Acad. Sci. USA* **92**: 3636-3641.

163. Tachida, H. and Iizuka, M. (1992). Persistence of repeated sequences that evolve by replication slippage. *Genetics* **131**: 471-478.

164. Takenawa, T. and Suetsugu, S. (2007). The WASP-WAVE protein network: connecting the membrane to the cytoskeleton. *Nat. Rev. Mol. Cell. Biol.* **8(1)**: 37-48.

165. Tatrai, P., Dudas, J., Batmunkh, E., Mathe, M., Zalatnai, A., Schaff, Z., Ramadori, G. and Kovalszky, I, (2006). Agrin, a novel basement membrane component in human and rat liver, accumulates in cirrhosis and hepatocellular carcinoma. *J. Canc. Res. Clin. Oncol.* **(2)**: 80-86.

166. Tautz, D. (1989). Hyper-variability of simple sequences as a general source for polymorphic DNA markers. *Nucl. Acids Res.* **17**: 6463-6471.

167. Templeton, A.R., Clark, A.G., Weiss, K.M., Nickerson, D.A., Boerwinkle, E, and Sing, C.F. (2000). Recombinational and mutational hot spots within the human lipoprotein lipase gene. *Am. J. Hum. Genet.* **66**: 69-83.

168. Termine, J.D., Kleinman, H.K., Whitson, S.W., Conn, K.M., McGarvey, M.L. and Martin, G.R. (1981). Osteonectin, a bone-specific protein linking mineral to collagen. *Cell* **26**: 99-105.

169. Thommes, K., Lennartsson, J., Carlberg, M. and Ronnstrand, L. (1999). Identification of Tyr-703 and Tyr-936 as the primary association sites for Grb2 and Grb7 in the *c-kit*/stem cell factor receptor. *Biochem. J.* **341**: 211–216.

170. Tian, B., White, R.J., Xia, T., Welle, S., Turner, D.H., Mathews, M.B., and Thornton, C.A. (2000). Expanded CUG repeats RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* **6**: 79–87.

171. Timchenko, N. A., Welm, A.L., Lu, X. and Timchenko, L.T. (1999). CUG repeat binding protein (CUGBP1) interacts with the 59 region of C/EBP-beta mRNA and regulates translation of C/EBP-beta isoforms. *Nucl. Acids Res.* **7**:4517–4525.

172. Timple, R. (1996). Macromolecular organization of basement membranes. *Curr. Opin. Cell. Biol.* **8**: 618-624.

173. Ting, C.H. Rosemberg, M.P., Snow, C.M., Samuelson, M.C. and Meisler, M.H. (1992). Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev.* **6**:1457-1465.

174. Toth, G., Gaspari, Z. and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10**: 967-981.

175. Toutenhoofd, S.L., Garcia, F., Zacharias, D.A., Wilson, R.A. and Strehler, E.E. (1998). Minimum CAG repeat in the human calmodulin-1 gene 5' untranslated region is required for full expression. *Biochim. Biophys. Acta.* **1398**:315–320.

176. Trapitz, P., Wlaschek, M. and Bunemann, H. (1988). Structure and function of Y chromosomal DNA. II. Analysis of lampbrush loop associated transcripts in nuclei of primary spermatocytes of *Drosophila hydei* by *in situ* hybridization using asymetric RNA probes of four different families of repetitive DNA. *Chromosoma* **96**: 159–170.

177. Treco, D. and Arnheim, N. (1986). The evolutionary conserved repetitive sequence d(TGáAC)n promotes reciprocal exchange and

generates unusual recombinant tetrads during yeast meiosis. *Mol. Cell. Biol.* **6**: 3934-3947.

178. Ugarkovic, D. (1995). Functional elements residing within satellite DNA's. *EMBO rep.* **6(11)**:1035-1039.
179. Ugarkovic, D. and Plohl, M. (2002). Variation in satellite DNA profiles-causes and effects. *EMBO J.* **21**: 5955–5959.
180. van Belkum, A., xScherer, L. van Alphen, S. and Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.* **62**:275–293.
181. Vannahme, C., Smyth, N., Miosge, N., Gosling, S., Frie, C., Paulsson, M., Maurer, P. and Hartmann, U. (2002). Characterization of *SMOC*-1, a novel modular calcium-binding protein in basement membranes. *J. Biol. Chem.* **277**: 37977–37986.
182. Varley, J.M., Macgregor, H.C., Nardi, I., Andrews, C., Erba, H.P. (1980). Cytological evidence of transcription of highly repeated DNA sequence during the lampbrush stage in *Triturus*. *Chromosoma* **100**: 15–31.
183. Verdel, A., Jia, S., Gerber, S., Suglyama, T., Gygi, S., Grewal, S.I., Moazed, D. (2004). RNAi-mediated targeting of heterochromatin with the RITS complex. *Science* **303**: 672–676.
184. Vergnaud, G., Denoeud, F. (2000). Minisatellites: Mutability and Genome architecture. *Genome Res.* **10**: 899-907.
185. Verstrepen, K.J., Jansen, A., Lewitter, F., Fink, G.R. (2005). Intragenic tandem repeats generated functional variability. *Nat. Genet.* **37(9)**: 986-990.
186. Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I.S., Martienssen, R.A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science* **297**: 1833–1837.
187. Wahls, W.P. and Moore, P.D. (1990a) Relative frequencies of homologous recombination between plasmids introduced into DNA repair deficient and other mammalian somatic cell lines. *Soma. Cell Mol. Genet.* 16: 321-329.

188. Wahls, W.P. and Moore, P.D. (1990b). Homologous recombination enhancement conferred by the Z-DNA motif d(TG)30 is abrogated by simian virus 40 T antigen binding to adjacent DNA sequences. *Mol. Cell. Biol.* **10**: 794-800.

189. Wang, Z., Weber, J.L., Zhong, G. and Tanksley, S.D. (1994). Survey of plant short tandem DNA repeats. *Theor. Appl. Genet.* **88**: 1-6.

190. Wang, A.H., Gregoire, S., Zika, E., Xiao, L., Li, C.S., Li, H., Wright, K.L., Ting, J.P. and Yang, X.J. (2005). Identification of the ankyrin repeat proteins ANKRA and RFXANK as novel partners of class IIa histone deacetylases. *J. Biol. Chem.* **280(32)**: 29117-29127.

191. Weitzel, J.N., Hows, J.M., Jeffreys, A.J., Min, G.L. and Goldman, J.M. (1988). Use of a hypervariable minisatellite DNA probe (33.15) for evaluating engraftment two or more years after bone marrow transplantation for aplastic anaemia. *Br. J. Haematol.* **70(1)**: 91-97.

192. Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl. Acids Res.* **18**: 6431–6535.

193. Winter, E. and Varshavsky, A. (1989). A DNA binding protein that recognizes oligo(dA) Ð oligo(dT) tracts. *EMBO J.* **8**: 1867Ð 1877.

194. Wren, J.D., Forgacs, E., Fondon, J.W. Pertsemlidis, A., Cheng, S.Y., Gallardo, T., Williams, R.S., Shohet, R.V., Minna, J.D. and Garner, H.R. (2000). Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am. J. Hum. Genet.* **67**:345–356.

195. Wrobel, K.H., Mademann, R. and Sinowatz, F. (1979). The lamina propria of the bovine seminiferous tubule. *Cell Tissue Res.* **202(3)**: 357-377.

196. Wykes, S.M., Visscher, D.W., Krawetz, S.A. (1997). Haploid transcripts persist in mature human spermatozoa. *Mol. Hum. Reprod.* **3(1)**: 15-19.

197. Young, E.T., Sloan, J.S. and van Riper, K. (2000). Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics* 154: 1053-1068.

198. Zeisberg, M., Kramer, K., Sindhi, N., Sarkar, P., Upton, M. and Kalluri, R. (2006). De-differentiation of primary human hepatocytes depends on the composition of specialized liver basement membrane. *Mol. Cell. Biochem.* **283(1-2)**: 181-189.

PUBLICATIONS

8. LIST OF PUBLICATIONS FROM THIS STUDY

I. **Jyoti Srivastava**, Sanjay Premi, Sudhir Kumar and Sher Ali (2008) "Organization and differential expression of GACA/GATA tagged Somatic and Spermatozoal transcriptomes in buffalo *Bubalus bubalis*" *BMC Genomics,* **In Press**, 000-000.

II. **Jyoti Srivastava**, Sanjay Premi, Sudhir Kumar, Iqbal Parwez and Sher Ali (2007) "Characterization of *Smoc*-1 uncovers two transcript variants showing differential tissue and age specific expression in *Bubalus bubalis*" *BMC Genomics,* **8,** 436.

III. **Jyoti Srivastava,** Sanjay Premi, Lalit C. Garg, and Sher Ali (2006) "Organizational and Expressional Uniqueness of a Testis Specific mRNA transcript of Protooncogene *c-kit* Receptor in Water Buffalo *Bubalus bubalis", DNA Cell Biol.,* **25,** 501-513.

IV. **Jyoti Srivastava**, Sanjay Premi, Deepali Pathak, Sudhir Kumar and Sher Ali (2006) "Development of Molecular Markers for Characterization of Stem Cell Lineages", *Pro. Ind. Nat. Sci. Acad. (PINSA)*, **72,** 83-89.

V. Deepali Pathak, **Jyoti Srivastava**, Sanjay Premi, Madhulika Tiwari, Zaid Ahsan, Lalit C. Garg, Sudhir Kumar and Sher Ali (2006) "Chromosomal Localization, Copy Number Assessment and Transcriptional Status of *Bam*HI Repeat Fractions in Water Buffalo *Bubalus bubalis", DNA Cell Biol.,* **25(4),** 206-214.

VI. **Jyoti Srivastava**, Sanjay Premi, Deepali Pathak, Zaid Ahsan, Madhulika Tiwari, Lalit C. Garg and *Sher Ali (2006) "Transcriptional Status of Known and Novel Genes Tagged with Consensus of 33.15 Repeat Loci Employing Minisatellite Associated Sequence Amplification (MASA) and Real Time PCR in Water Buffalo *Bubalus bubalis", DNA Cell Biol.,* **25(1),** 31-48.

BMC Genomics

BioMed Central

Research article

Open Access

Characterization of *Smoc-1* uncovers two transcript variants showing differential tissue and age specific expression in *Bubalus bubalis*

Jyoti Srivastava, Sanjay Premi, Sudhir Kumar, Iqbal Parwez and Sher Ali*

Address: Molecular Genetics Laboratory, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi-110 067, India

Email: Jyoti Srivastava - jayanshi@gmail.com; Sanjay Premi - sanjaypre@gmail.com; Sudhir Kumar - panwarsk@yahoo.com; Iqbal Parwez - iparwez2002@yahoo.co.in; Sher Ali* - alisher@nii.res.in

* Corresponding author

Published: 28 November 2007

BMC Genomics 2007, **8**:436 doi:10.1186/1471-2164-8-436

Received: 13 July 2007
Accepted: 28 November 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/436

Abstract

Background: Secreted modular calcium binding protein-1 (*Smoc-1*) belongs to the BM-40 family which has been implicated with tissue remodeling, angiogenesis and bone mineralization. Besides its anticipated role in embryogenesis, *Smoc-1* has been characterized only in a few mammalian species. We made use of the consensus sequence (5' CACCTCTCCACCTGCC 3') of 33.15 repeat loci to explore the buffalo transcriptome and uncovered the *Smoc-1* transcript tagged with this repeat. The main objective of this study was to gain an insight into its structural and functional organization, and expressional status of *Smoc-1* in water buffalo, *Bubalus bubalis*.

Results: We cloned and characterized the buffalo *Smoc-1*, including its copy number status, *in-vitro* protein expression, tissue & age specific transcription/translation, chromosomal mapping and localization to the basement membrane zone. Buffalo *Smoc-1* was found to encode a secreted matricellular glycoprotein containing two EF-hand calcium binding motifs homologous to that of BM-40/SPARC family. In buffalo, this single copy gene consisted of 12 exons and was mapped onto the acrocentric chromosome 11. Though this gene was found to be evolutionarily conserved, the buffalo *Smoc-1* showed conspicuous nucleotide/amino acid changes altering its secondary structure compared to that in other mammals. *In silico* analysis of the *Smoc-1* proposed its glycoprotein nature with a calcium dependent conformation. Further, we unveiled two transcript variants of this gene, varying in their 3'UTR lengths but both coding for identical protein(s). *Smoc-1* evinced highest expression of both the variants in liver and modest to negligible in other tissues. The relative expression of variant-02 was markedly higher compared to that of variant-01 in all the tissues examined. Moreover, expression of *Smoc-1*, though modest during the early ages, was conspicuously enhanced after 1 year and remained consistently higher during the entire life span of buffalo with gradual increment in expression of variant-02. Immunohistochemically, *Smoc-1* was localized in the basement membrane zones and extracellular matrices of various tissues.

Conclusion: These data added to our understandings about the tissue, age and species specific functions of the *Smoc-1*. It also enabled us to demonstrate varying expression of the two transcript variants of *Smoc-1* amongst different somatic tissues/gonads and ages, in spite of their identical coding frames. Pursuance of these variants for their roles in various disease phenotypes such as hepatocellular carcinoma and angiogenesis is envisaged to establish broader biological significance of this gene.

DNA AND CELL BIOLOGY
Volume 25, Number 9, 2006

Pp. 501–513

Organizational and Expressional Uniqueness of a Testis-Specific mRNA Transcript of Protooncogene *c-kit* Receptor in Water Buffalo *Bubalus bubalis*

JYOTI SRIVASTAVA,[1] SANJAY PREMI,[1] LALIT C. GARG,[2] and SHER ALI[1]

ABSTRACT

Protooncogene *c-kit* receptor is implicated with spermatogenesis, melanogenesis, and hematopoeisis, and undergoes tissue/stage specific alternate splicing. We have isolated 2973-bp full-length cDNA sequence (CDS) of this gene from testis and other tissues of water buffalo *Bubalus bubalis*. Upon comparison, the *c-kit* sequences showed tissue specific nucleotide changes resulting in novel truncated peptides. These peptides lacked intracellular and/or transmembrane domains in all the tissues except testis. Other alternately spliced tissue-specific transcripts were also detected, which are the integral parts of the open reading frame and have been reported in other mammals. Phylogenetic analysis of the sequences revealed unique tyrosine kinase domain in buffalo. Copy number calculation and expressional analysis of *c-kit* using real-time PCR established its single copy status and highest expression (137–177 folds) in testis compared to that (least) in liver. *c-kit* expression was detected in semen samples although 10 times lesser compared to that in testis. The highest expression of *c-kit* in testis and the presence of mRNA transcript in sperms substantiate its predominant role in spermatogenesis. This study establishes unequivocal involvement of an autosomal gene *c-kit* receptor in testicular function.

INTRODUCTION

THE PLEIOTROPIC PROTOONCOGENE *c-kit* receptor belongs to transmembrane receptor tyrosine kinases (RTK) family type-3 (Andre *et al.*, 1992; Chabot *et al.*, 1988) similar to the receptors for platelet derived growth factor (PDGF) and macrophage-colony-stimulating factor (M-CSF) (Qiu *et al.*, 1988). The *c-kit* glycoprotein includes an immunoglobulin like extracellular, a single transmembrane, and an intracellular tyrosine-kinase domain (Galli *et al.*, 1994; Gokkel *et al.*, 1992). The cytoplasmic domain is divided by a hydrophilic kinase insert into Adenosine triphosphate (ATP) binding and phosphotransferase regions (Hashimoto *et al.*, 2003). The exons and exon–intron boundary regions of this gene are conserved, but extracellular domain and intronic sequences show variations across the species (Reith *et al.*, 1991; Crosier *et al*, 1993). Despite organizational variations within the extracellular domain, the *c-kit* gene is functionally conserved and undergoes tissue and stage-specific alternative splicing (Serve *et al.*, 1995). Two isoforms of *c-kit* in the mouse and four in humans with alteration of four amino acids GNNK in the juxtamembrane region have been reported (Reith *et al.*, 1991; Crosier *et al*, 1993; Serve *et al.*, 1995). In the presence of a stem cell factor (SCF), the GNNK form induces anchorage-independent growth, loss of contact inhibition, and tumorigenecity (Voytuyk *et al.*, 2003). This gene is expressed in differentiating Spermatogonia A (Yoshimaga *et al.*, 1991; Schrans-Stassen *et al.*, 1999) and B (Manova *et al.*, 1990), in pre-meiotic (Manova *et al.*, 1993) and meiotic spermatocytes (Vincent *et al.*, 1998). However, round spermatids express an alternative messenger driven by activation of cell and stage-specific promoters in the 16th intron (Sorrentino *et al.*, 1991; Albanesi *et al.*, 1996), which encodes for a truncated *c-kit* protein (tr-kit). The tr-kit lacks the extracellular domain, transmembrane domain, and ATP-binding sites in the intracellular region, and encodes short sequences of the interkinase segment, phosphotransferase domain, and the carboxyl-terminal tail of the receptor. Similarly, expression of SCF has been documented in Sertoli cells during different stages of development both as soluble and membrane-bound proteins (Anderson *et al.*, 1990; Toksoz *et al.*, 1990).

[1]Molecular Genetics Laboratory, [2]Gene Regulation Laboratory, National Institute of Immunology, New Delhi, India.

DNA AND CELL BIOLOGY
Volume 25, Number 4, 2006

Pp. 206–214

Chromosomal Localization, Copy Number Assessment, and Transcriptional Status of *Bam*HI Repeat Fractions in Water Buffalo *Bubalus bubalis*

DEEPALI PATHAK,[1] JYOTI SRIVASTAVA,[1] SANJAY PREMI,[1] MADHULIKA TIWARI,[2] LALIT C. GARG,[2] SUDHIR KUMAR,[1] and SHER ALI[1]

ABSTRACT

Higher eukaryotes contain a wide variety of repetitive DNA, although their functions often remain unknown. We describe cloning, chromosomal localization, copy number assessment, and transcriptional status of 1378- and 673-bp repeat fractions in the buffalo genome. The pDS5, representing the 1378-bp fragment, showed FISH signals in the centromeric region of acrocentric chromosomes only, whereas pDS4, corresponding to 673 bp, detected signals in the centromeric regions of all the chromosomes. Crosshybridization studies of pDS5 and pDS4 with genomic DNA from different sources showed signals only in buffalo, cattle, goat, and sheep. Real-time PCR analysis uncovered 1234 and 3420 copies of pDS5 and pDS4 fragments per the haploid genome, corresponding to 30 and 68 copies per chromosome, respectively. Analysis of cDNA from different tissues of buffalo with Real-time PCR showed maximum expression of pDS5 and pDS4 in the spleen and liver, respectively. Phylogenetic analysis of these sequences showed a close relationship between buffalo and cattle. The prospect of this approach in comparative genomics is highlighted.

INTRODUCTION

SATELLITE DNA REPRESENTS tandemly repeated sequences, organized in long, usually megabase arrays, and located in the pericentromeric and/or telomeric heterochromatic regions (Charlesworth *et al.*, 1994). Nucleotide changes and copy number variations fuel the process of their evolution within and across the species (Ugarkovic and Plohl, 2002). Satellite fraction(s), although not conserved evolutionarily (Amor and Choo, 2002), are unique to a species and usually show similarity among related groups of animals (Ali and Gangadharan, 2000; Henikoff *et al.*, 2001). With respect to the functional role of these sequences, uncertainty persisted for a long time, and it was largely believed that they represent detritus part of the genome (Ohno, 1972). However, recent studies have shown repeat elements influencing the structure, function, and evolution of the chromosomes in the host species (Sinden, 1999; Dey and Rath, 2005). Studies on centromeric and telomeric sequences, retrotransposons, and Alu-repeats have substantiated this view (Grady *et al.*, 1992; Wolffe, 1989). Similarly, expansion of trinucleotide repeats leading to hereditary neurodegenerative diseases in humans has highlighted the importance of repeat elements in mammalian genome (Paulson, 1999). Short tandem repeat (STR) motifs and microsatellites are frequently used as markers for genotyping, genome mapping, species diversity, and molecular mining of the satellite tagged transcribing sequences (Chattopadhyay *et al.*, 2001; Srivastava *et al.*, 2006). Centromeric heterochromatic sequences participate in the kinetochore activities during cell division (Mellone *et al.*, 2003). These sequences have been well characterized in humans (Schueler *et al.*, 2001), the mouse (Broccoli *et al.*, 1991), *Drosophila* (Sun *et al.*, 1997), and cattle (Pluciennicczak *et al.*, 1982; Taparowsky and Gerbi, 1982; Nijman and Lenstra, 2001). However, their organizational, evolutionary, and transcriptional status in buffalo genome remains unknown, despite the fact that this is an important species for agricultural and dairy industries throughout the Indian subcontinent.

We describe cloning, chromosomal localization by fluorescence *in situ* hybridization (FISH), copy number assessment, and transcriptional status of 1378 and 673 bp repeat fractions

[1]Molecular Genetics Laboratory, [2]Gene Regulation Laboratory, National Institute of Immunology, New Delhi, India.

D A A D CELL BIOLOGY
Volume 25, umbe r 1, 2006

Pp. 31–48

Transcriptional Status of Known and Novel Genes Tagged with Consensus of 33.15 Repeat Loci Employing Minisatellite-Associated Sequence Amplification (MASA) and Real-Time PCR in Water Buffalo, *Bubalus bubalis*

JYOTI SRIVASTAVA,[1] SANJAY PREMI,[1] DEEPALI PATHAK,[1] ZAID AHSAN,[2] MADHULIKA TIWARI,[2] LALIT C. GARG,[2] and SHER ALI[1]

ABSTRACT

We conducted minisatellite-associated sequence amplification (MASA) with an oligo (5′ CACCTCTCCACCTGCC 3′) based on consensus of 33.15 repeat loci using cDNA from the testis, ovary, spleen, kidney, heart, liver, and lung of water buffalo *Bubalus bubalis* and uncovered 25 amplicons of six different sizes (1263, 846/847, 602, 576, 487, and 324 base pairs). These fragments, cloned and sequenced, were found to represent several functional, regulatory, and structural genes. Blast search of all the 25 amplicons showed homologies with 43 transcribing genes across the species. Of these, the 846/847-bp fragment, having homology with the adenylate kinase gene, showed nucleotide changes at six identical places in the ovary and testis. The 1263; 324; and 487-bp fragments showed homology with the secreted modular calcium binding protein (SMOC-1), leucine-rich repeat neuronal 6A (LRRN6A) mRNA, and human TTTY5 mRNA, respectively. Real-time PCR showed maximum expression of AKL, LRRN6A, and T-cell receptor gamma (TCR-γ)-like genes in the testis, SMOC-1 in the liver, and the T-cell receptor-like (TCRL) gene in the spleen compared to those used as endogenous control. We construe that these genes have evolved from a common progenitor and conformed to various biological functions during the course of evolution. MASA approach coupled with real-time PCR has potentials to uncover accurate expression of a large number of genes within and across the species circumventing the screening of cDNA library.

INTRODUCTION

An eukaryotic genome contains a sizable portion of repetitive DNA besides single or multiple copies of the transcribing sequence (Nadir *et al.*, 1996; Wickstead *et al.*, 2004). Coding sequences may be organized in the proximity of the noncoding short tandem repeats (STR), or may harbor such motifs within themselves (Johansson *et al.*, 1992). A large number of mega-, mini-, and microsatellite sequences have been characterized from a number of species (Jeffreys *et al.*, 1988). Some of these are evolutionarily conserved (Tautz, 1989; Robles *et al.*, 2004), whereas others remain unique to a given genome (Ali *et al.*, 1999). Repeat sequences are known to shrink and expand, fuelling the process of copy number alteration (John and Ali, 1997; Nakamura *et al.*, 1987), and have been associated with tumorigenesis and genetic anomalies (Epplen, 1988; Kizawa *et al.*, 2005; Ross *et al.*, 2005). The 16 nucleotide long (5′ CACCTCTCCACCTGCC 3′) consensus of 33.15 repeat loci originating from the human myoglobin gene (7q35–q36) studied in a number of species (Ali and Wallace, 1988; Jeffreys *et al.*, 1985; Weitzel *et al.*, 1988) have also been found to be associated with heterochromatic sequences of the human Y chromosome (Bashamboo and Ali, 2001). We wanted to ascertain if this repeat motif is part of mRNA transcripts of structural, functional, and regulatory genes and involved in up- or downregulation of these genes in somatic tissues and gonads.

[1]Molecular Genetics Laboratory, [2]Gene Regulation Laboratory, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi, India.